**Supplemental File 1.**

**Title:**

A snapshot of 3,649 web-based services published between 1994 and 2017 shows a decrease in availability after two years

**Authors:**

Ágnes Ősz, Lőrinc Sándor Pongor, Danuta Szirmai, Balázs Győrffy

**Description:**

Parser input file

**Usage:**

Pubmed abstract file based on searching using of following terms: “www[Title/Abstract] OR http[Title/Abstract]) AND (online tool[Title/Abstract] OR web server[Title/Abstract] OR server[Title/Abstract]”

====================================================================

1. Clin Orthop Relat Res. 2017 Apr;475(4):1252-1261. doi: 10.1007/s11999-016-5187-3.

Epub 2016 Dec 1.

Can We Estimate Short- and Intermediate-term Survival in Patients Undergoing

Surgery for Metastatic Bone Disease?

Forsberg JA(1,)(2,)(3), Wedin R(4), Boland PJ(5), Healey JH(5).

Author information:

(1)Regenerative Medicine, Naval Medical Research Center, 503 Robert Grant Avenue,

Silver Spring, MD, 20910, USA. jaforsberg@me.com. (2)Uniformed Services

University-Walter Reed Department of Surgery, Bethesda, MD, USA.

jaforsberg@me.com. (3)Section of Orthopaedics and Sports Medicine, Department of

Molecular Medicine and Surgery, Karolinska Institute, Karolinska University

Hospital, Stockholm, Sweden. jaforsberg@me.com. (4)Section of Orthopaedics and

Sports Medicine, Department of Molecular Medicine and Surgery, Karolinska

Institute, Karolinska University Hospital, Stockholm, Sweden. (5)Orthopaedic

Surgery Service, Department of Surgery, Memorial Sloan-Kettering Cancer Center

and Weill College of Medicine, Cornell University, New York, NY, USA.

BACKGROUND: Objective means of estimating survival can be used to guide surgical

decision-making and to risk-stratify patients for clinical trials. Although a

free, online tool ( www.pathfx.org ) can estimate 3- and 12-month survival,

recent work, including a survey of the Musculoskeletal Tumor Society, indicated

that estimates at 1 and 6 months after surgery also would be helpful. Longer

estimates help justify the need for more durable and expensive reconstructive

options, and very short estimates could help identify those who will not survive

1 month and should not undergo surgery. Thereby, an important use of this tool

would be to help avoid unsuccessful and expensive surgery during the last month

of life.

QUESTIONS/PURPOSES: We seek to provide a reliable, objective means of estimating

survival in patients with metastatic bone disease. After generating models to

derive 1- and 6-month survival estimates, we determined suitability for clinical

use by applying receiver operator characteristic (ROC) (area under the curve

[AUC] > 0.7) and decision curve analysis (DCA), which determines whether using

PATHFx can improve outcomes, but also discerns in which kinds of patients PATHFx

should not be used.

METHODS: We used two, existing, skeletal metastasis registries chosen for their

quality and availability. Data from Memorial Sloan-Kettering Cancer Center

(training set, n = 189) was used to develop two Bayesian Belief Networks trained

to estimate the likelihood of survival at 1 and 6 months after surgery. Next,

data from eight major referral centers across Scandinavia (n = 815) served as the

external validation set-that is, as a means to test model performance in a

different patient population. The diversity of the data between the training set

from Memorial Sloan-Kettering Cancer Center and the Scandinavian external

validation set is important to help ensure the models are applicable to patients

in various settings with differing demographics and treatment philosophies. We

considered disease-specific, laboratory, and demographic information, and the

surgeon's estimate of survival. For each model, we calculated the area under the

ROC curve (AUC) as a metric of discriminatory ability and the Net Benefit using

DCA to determine whether the models were suitable for clinical use.

RESULTS: On external validation, the AUC for the 1- and 6-month models were 0.76

(95% CI, 0.72-0.80) and 0.76 (95% CI, 0.73-0.79), respectively. The models

conferred a positive net benefit on DCA, indicating each could be used rather

than assume all patients or no patients would survive greater than 1 or 6 months,

respectively.

CONCLUSIONS: Decision analysis confirms that the 1- and 6-month Bayesian models

are suitable for clinical use.

CLINICAL RELEVANCE: These data support upgrading www.pathfx.org with the

algorithms described above, which is designed to guide surgical decision-making,

and function as a risk stratification method in support of clinical trials. This

updating has been done, so now surgeons may use any web browser to generate

survival estimates at 1, 3, 6, and 12 months after surgery, at no cost. Just as

short estimates of survival help justify palliative therapy or less-invasive

approaches to stabilization, more favorable survival estimates at 6 or 12 months

are used to justify more durable, complicated, and expensive reconstructive

options.

DOI: 10.1007/s11999-016-5187-3

PMID: 27909972

2. J Comput Biol. 2017 Mar;24(3):255-265. doi: 10.1089/cmb.2016.0074. Epub 2016 Aug

5.

NIAS-Server: Neighbors Influence of Amino acids and Secondary Structures in

Proteins.

Borguesan B(1), Inostroza-Ponta M(2), Dorn M(1).

Author information:

(1)1 Institute of Informatics, Federal University of Rio Grande do Sul , Porto

Alegre, Brazil . (2)2 Departamento de Ingeniería Informática, Center for

Biotechnology and Bioengineering , Universidad de Santiago de Chile, Santiago,

Chile .

The exponential growth in the number of experimentally determined

three-dimensional protein structures provide a new and relevant knowledge about

the conformation of amino acids in proteins. Only a few of probability densities

of amino acids are publicly available for use in structure validation and

prediction methods. NIAS (Neighbors Influence of Amino acids and Secondary

structures) is a web-based tool used to extract information about conformational

preferences of amino acid residues and secondary structures in

experimental-determined protein templates. This information is useful, for

example, to characterize folds and local motifs in proteins, molecular folding,

and can help the solution of complex problems such as protein structure

prediction, protein design, among others. The NIAS-Server and supplementary data

are available at http://sbcb.inf.ufrgs.br/nias .

DOI: 10.1089/cmb.2016.0074

PMID: 27494258

3. Gene. 2017 Feb 20;602:1-7. doi: 10.1016/j.gene.2016.11.021. Epub 2016 Nov 11.

Gly-PseAAC: Identifying protein lysine glycation through sequences.

Xu Y(1), Li L(2), Ding J(3), Wu LY(4), Mai G(5), Zhou F(6).

Author information:

(1)Department of Information and Computer Science, University of Science and

Technology Beijing, Beijing 100083, China. Electronic address: xuyan@ustb.edu.cn.

(2)Department of Information and Computer Science, University of Science and

Technology Beijing, Beijing 100083, China. Electronic address:

s20157158@ustb.edu.cn. (3)Department of Information and Computer Science,

University of Science and Technology Beijing, Beijing 100083, China. Electronic

address: artin@ustb.edu.cn. (4)Institute of Applied Mathematics, Academy of

Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190,

China. Electronic address: wulingyun@gmail.com. (5)Center for Synthetic Biology

Engineering Research, Shenzhen Institutes of Advanced Technology, Chinese Academy

of Sciences, Shenzhen, Guangdong 518055, China. Electronic address:

gq.mai@siat.ac.cn. (6)College of Computer Science and Technology, Jilin

University, Changchun 130012, China; Key Laboratory of Symbolic Computation and

Knowledge Engineering of Ministry of Education, Jilin University, Changchun,

Jilin 130012, China.. Electronic address: ffzhou@jlu.edu.cn.

BACKGROUND: Similar to the regular enzymatic glycosylation, glycation also

attaches a sugar molecule to a peptide, but does not need the help of an enzyme.

Glycation may occur both inside and outside the host body, and will compete with

the glycosylation procedure for functional regulation of mature protein products.

The glycated residues do not show significant patterns, which make both in silico

sequence-level predictors and wet-lab validations a major challenge. This study

hypothesizes that a better feature set formulated from the glycated flanking

peptides may lead to a good glycation prediction program.

RESULTS: We explored the application of sequence order information and position

specific amino acid propensity (PSAAP) in the glycation residue prediction

problem. The PSAAP demonstrated its ability to discriminate the glycated residues

from the background control peptides. A Support Vector Machine (SVM) model was

constructed from the training dataset and achieved 68.91% in the overall

accuracy. The model also achieves 0.7258 and 0.3198 in the Area under the ROC and

Matthew's Correlation Coefficient, respectively. The user-friendly online version

of the proposed algorithm may be found on the web server Gly-PseAAC at

http://app.aporc.org/Gly-PseAAC/.

CONCLUSION: The feature set PSAAP was calculated and led to a useful

classification of glycation residues.

Copyright © 2016 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.gene.2016.11.021

PMID: 27845204 [Indexed for MEDLINE]

4. J Mol Biol. 2017 Feb 3;429(3):365-371. doi: 10.1016/j.jmb.2016.12.004. Epub 2016

Dec 10.

Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions

in Protein Structures.

Jubb HC(1), Higueruelo AP(2), Ochoa-Montaño B(2), Pitt WR(3), Ascher DB(4),

Blundell TL(5).

Author information:

(1)Department of Biochemistry, Sanger Building, University of Cambridge, 80

Tennis Court Road, Cambridge CB2 1GA, UK. Electronic address: hj4@sanger.ac.uk.

(2)Department of Biochemistry, Sanger Building, University of Cambridge, 80

Tennis Court Road, Cambridge CB2 1GA, UK. (3)UCB, 208 Bath Road, Slough, West

Berkshire SL1 3WE, UK. (4)Department of Biochemistry, Sanger Building, University

of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK. Electronic address:

david.ascher@unimelb.edu.au. (5)Department of Biochemistry, Sanger Building,

University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK. Electronic

address: tlb20@cam.ac.uk.

Interactions between proteins and their ligands, such as small molecules, other

proteins, and DNA, depend on specific interatomic interactions that can be

classified on the basis of atom type and distance and angle constraints.

Visualisation of these interactions provides insights into the nature of

molecular recognition events and has practical uses in guiding drug design and

understanding the structural and functional impacts of mutations. We present

Arpeggio, a web server for calculating interactions within and between proteins

and protein, DNA, or small-molecule ligands, including van der Waals', ionic,

carbonyl, metal, hydrophobic, and halogen bond contacts, and hydrogen bonds and

specific atom-aromatic ring (cation-π, donor-π, halogen-π, and carbon-π) and

aromatic ring-aromatic ring (π-π) interactions, within user-submitted

macromolecule structures. PyMOL session files can be downloaded, allowing

high-quality publication images of the interactions to be generated. Arpeggio is

implemented in Python and available as a user-friendly web interface at

http://structure.bioc.cam.ac.uk/arpeggio/ and as a downloadable package at

https://bitbucket.org/harryjubb/arpeggio.

Crown Copyright © 2016. Published by Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jmb.2016.12.004

PMCID: PMC5282402

PMID: 27964945

5. J Mol Biol. 2017 Feb 3;429(3):382-389. doi: 10.1016/j.jmb.2016.11.034. Epub 2016

Dec 10.

iFrag: A Protein-Protein Interface Prediction Server Based on Sequence Fragments.

Garcia-Garcia J(1), Valls-Comamala V(2), Guney E(3), Andreu D(4), Muñoz FJ(2),

Fernandez-Fuentes N(5), Oliva B(6).

Author information:

(1)Structural Bioinformatics Laboratory, Department of Experimental and Health

Sciences, Universitat Pompeu Fabra, C/Doctor Aiguader 88, 08003 Barcelona, Spain.

(2)Laboratory of Molecular Physiology, Department of Experimental and Health

Sciences, Universitat Pompeu Fabra, C/Doctor Aiguader 88, 08003 Barcelona, Spain.

(3)Center for Complex Network Research and Department of Physics, Northeastern

University, Boston, 02115, MA, USA; Joint IRB-BSC-CRG Program in Computational

Biology, Institute for Research in Biomedicine (IRB Barcelona), C/Baldiri Reixac

10-12, 08028 Barcelona, Spain. (4)Laboratory of Proteomics and Protein Chemistry,

Department of Experimental and Health Sciences, Universitat Pompeu Fabra,

C/Doctor Aiguader 88, 08003 Barcelona, Spain. (5)Institute of Biological,

Environmental and Rural Sciences, Aberystwyth University, Gogerddan Campus, SY23

3EB Aberystwyth, UK. (6)Structural Bioinformatics Laboratory, Department of

Experimental and Health Sciences, Universitat Pompeu Fabra, C/Doctor Aiguader 88,

08003 Barcelona, Spain. Electronic address: baldo.oliva@upf.edu.

Protein-protein interactions (PPIs) are crucial in many biological processes. The

first step towards the molecular characterisation of PPIs implies the charting of

their interfaces, that is, the surfaces mediating the interaction. To this end,

we present here iFrag, a sequence-based computational method that infers possible

interacting regions between two proteins by searching minimal common sequence

fragments of the interacting protein pairs. By utilising the sequences of two

interacting proteins (queries), iFrag derives a two-dimensional matrix computing

a score for each pair of residues that relates to the presence of similar regions

in interolog protein pairs. The scoring matrix is represented as a heat map

reflecting the potential interface regions in both query proteins. Unlike

existing approaches, iFrag does not require three-dimensional structural

information or multiple sequence alignments and can even predict small

interaction sites consisting only of few residues. Thus, predicted interfaces

range from short fragments composed of few residues to domains of proteins,

depending on available information on PPIs, as we demonstrate in several

examples. Moreover, as a proof of concept, we include the experimental validation

on the successful prediction of a peptide competing with the aggregation of

β-amyloid in Alzheimer's disease. iFrag is freely accessible at

http://sbi.imim.es/iFrag.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jmb.2016.11.034

PMID: 27956148

6. J Mol Biol. 2017 Feb 3;429(3):416-425. doi: 10.1016/j.jmb.2016.10.013. Epub 2016

Oct 12.

PROFEAT Update: A Protein Features Web Server with Added Facility to Compute

Network Descriptors for Studying Omics-Derived Networks.

Zhang P(1), Tao L(2), Zeng X(1), Qin C(1), Chen SY(1), Zhu F(3), Yang SY(4), Li

ZR(5), Chen WP(6), Chen YZ(7).

Author information:

(1)Bioinformatics and Drug Design Group, Department of Pharmacy, National

University of Singapore, 117543, Singapore. (2)School of Medicine, Hangzhou

Normal University, Hangzhou 310012, P.R. China. (3)Innovative Drug Research

Centre and College of Chemistry and Chemical Engineering, Chongqing University,

Chongqing 401331, P.R. China. (4)State Key Laboratory of Biotherapy and Cancer

Center, West China Hospital, West China Medical School, Sichuan University,

Chengdu 610041, P.R. China. (5)College of Chemistry, Sichuan University, Chengdu

610064, P.R. China; Molecular Medicine Research Center, State Key Laboratory of

Biotherapy, West China Hospital, West China School of Medicine, Sichuan

University, Chengdu 610041, P.R. China. Electronic address: lizerong@scu.edu.cn.

(6)Key Lab of Agricultural Products Processing and Quality Control of Nanchang

City, Jiangxi Agricultural University, Nanchang 330045, P.R. China. Electronic

address: iaochen@163.com. (7)Bioinformatics and Drug Design Group, Department of

Pharmacy, National University of Singapore, 117543, Singapore. Electronic

address: phacyz@nus.edu.sg.

The studies of biological, disease, and pharmacological networks are facilitated

by the systems-level investigations using computational tools. In particular, the

network descriptors developed in other disciplines have found increasing

applications in the study of the protein, gene regulatory, metabolic, disease,

and drug-targeted networks. Facilities are provided by the public web servers for

computing network descriptors, but many descriptors are not covered, including

those used or useful for biological studies. We upgraded the PROFEAT web server

http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi for computing up to 329

network descriptors and protein-protein interaction descriptors. PROFEAT network

descriptors comprehensively describe the topological and connectivity

characteristics of unweighted (uniform binding constants and molecular levels),

edge-weighted (varying binding constants), node-weighted (varying molecular

levels), edge-node-weighted (varying binding constants and molecular levels), and

directed (oriented processes) networks. The usefulness of the network descriptors

is illustrated by the literature-reported studies of the biological networks

derived from the genome, interactome, transcriptome, metabolome, and diseasome

profiles.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jmb.2016.10.013

PMID: 27742592

7. J Mol Biol. 2017 Feb 3;429(3):390-398. doi: 10.1016/j.jmb.2016.09.005. Epub 2016

Sep 10.

TMDOCK: An Energy-Based Method for Modeling α-Helical Dimers in Membranes.

Lomize AL(1), Pogozheva ID(2).

Author information:

(1)Department of Medicinal Chemistry, College of Pharmacy, University of

Michigan, Ann Arbor, MI 48109-1065, USA. Electronic address: almz@umich.edu.

(2)Department of Medicinal Chemistry, College of Pharmacy, University of

Michigan, Ann Arbor, MI 48109-1065, USA.

TMDOCK is a novel computational method for the modeling of parallel homodimers

formed by transmembrane (TM) α-helices. Three-dimensional (3D) models of dimers

are generated by threading a target amino acid sequence through several

structural templates, followed by local energy minimization. This is the first

method that identifies helix dimerization modes and ranks them based on the

calculated free energy of α-helix association. Free energy components include van

der Waals, hydrogen bonding, and dipole interactions; side-chain conformational

entropy; and solvation energy in the anisotropic lipid environment. TMDOCK

reproduced 26 experimental dimeric structures formed by TM α-helices of 21

single-pass membrane proteins (including 4 mutants) with Cα atom rmsd from 1.0 to

3.3Å. Assessment of dimerization heterogeneity of these TM domains demonstrated

that 7 of them have a unique dimer structure, 12 have at least 2 alternative

conformations, and 2 have a large number of different association modes. All

unique experimental structures of proteins from the first group and eight

structures from the second group were reproduced in computations as top-ranked

models. A fast version of the method is available through the web server

(http://membranome.org/tm\_server.php).

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jmb.2016.09.005

PMID: 27622289

8. Cogn Neurodyn. 2017 Feb;11(1):113-116. doi: 10.1007/s11571-016-9407-z. Epub 2016

Sep 15.

Parameterizable consensus connectomes from the Human Connectome Project: the

Budapest Reference Connectome Server v3.0.

Szalkai B(1), Kerepesi C(2), Varga B(1), Grolmusz V(3).

Author information:

(1)0000 0001 2294 6276grid.5591.8PIT Bioinformatics Group, Eötvös University,

Budapest, 1117 Hungary. (2)0000 0001 2294 6276grid.5591.8PIT Bioinformatics

Group, Eötvös University, Budapest, 1117 Hungary ; 0000 0001 2149

4407grid.5018.cInstitute for Computer Science and Control, Hungarian Academy of

Science, Budapest, 1111 Hungary. (3)0000 0001 2294 6276grid.5591.8PIT

Bioinformatics Group, Eötvös University, Budapest, 1117 Hungary ; Uratim Ltd.,

Budapest, 1118 Hungary.

Connections of the living human brain, on a macroscopic scale, can be mapped by a

diffusion MR imaging based workflow. Since the same anatomic regions can be

corresponded between distinct brains, one can compare the presence or the absence

of the edges, connecting the very same two anatomic regions, among multiple

cortices. Previously, we have constructed the consensus braingraphs on 1015

vertices first in five, then in 96 subjects in the Budapest Reference Connectome

Server v1.0 and v2.0, respectively. Here we report the construction of the

version 3.0 of the server, generating the common edges of the connectomes of

variously parameterizable subsets of the 1015-vertex connectomes of 477 subjects

of the Human Connectome Project's 500-subject release. The consensus connectomes

are downloadable in CSV and GraphML formats, and they are also visualized on the

server's page. The consensus connectomes of the server can be considered as the

"average, healthy" human connectome since all of their connections are present in

at least k subjects, where the default value of [Formula: see text], but it can

also be modified freely at the web server. The webserver is available at

http://connectome.pitgroup.org.

DOI: 10.1007/s11571-016-9407-z

PMCID: PMC5264751 [Available on 2018-02-01]

PMID: 28174617

9. J Comput Aided Mol Des. 2017 Feb;31(2):237-244. doi: 10.1007/s10822-016-9999-8.

Epub 2016 Dec 27.

Fast H-DROP: A thirty times accelerated version of H-DROP for interactive

SVM-based prediction of helical domain linkers.

Richa T(1), Ide S(1), Suzuki R(1), Ebina T(1,)(2), Kuroda Y(3).

Author information:

(1)Department of Biotechnology and Life Science, Tokyo University of Agriculture

and Technology, 12-24-16 Nakamachi, Koganei-shi, Tokyo, 184-8588, Japan.

(2)Department of Physiology, Graduate school of Medicine, The University of

Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan. (3)Department of

Biotechnology and Life Science, Tokyo University of Agriculture and Technology,

12-24-16 Nakamachi, Koganei-shi, Tokyo, 184-8588, Japan. ykuroda@cc.tuat.ac.jp.

Efficient and rapid prediction of domain regions from amino acid sequence

information alone is often required for swift structural and functional

characterization of large multi-domain proteins. Here we introduce Fast H-DROP, a

thirty times accelerated version of our previously reported H-DROP (Helical

Domain linker pRediction using OPtimal features), which is unique in specifically

predicting helical domain linkers (boundaries). Fast H-DROP, analogously to

H-DROP, uses optimum features selected from a set of 3000 ones by combining a

random forest and a stepwise feature selection protocol. We reduced the

computational time from 8.5 min per sequence in H-DROP to 14 s per sequence in

Fast H-DROP on an 8 Xeon processor Linux server by using SWISS-PROT instead of

Genbank non-redundant (nr) database for generating the PSSMs. The sensitivity and

precision of Fast H-DROP assessed by cross-validation were 33.7 and 36.2%, which

were merely ~2% lower than that of H-DROP. The reduced computational time of Fast

H-DROP, without affecting prediction performances, makes it more interactive and

user-friendly. Fast H-DROP and H-DROP are freely available from

http://domserv.lab.tuat.ac.jp/ .

DOI: 10.1007/s10822-016-9999-8

PMID: 28028736

10. Nucleic Acids Res. 2017 Jan 25;45(2):886-893. doi: 10.1093/nar/gkw749. Epub 2016

Aug 29.

Computational prediction of regulatory, premature transcription termination in

bacteria.

Millman A(1), Dar D(1), Shamir M(1), Sorek R(2).

Author information:

(1)Department of Molecular Genetics, Weizmann Institute of Science, Rehovot

76100, Israel. (2)Department of Molecular Genetics, Weizmann Institute of

Science, Rehovot 76100, Israel rotem.sorek@weizmann.ac.il.

A common strategy for regulation of gene expression in bacteria is conditional

transcription termination. This strategy is frequently employed by 5'UTR

cis-acting RNA elements (riboregulators), including riboswitches and attenuators.

Such riboregulators can assume two mutually exclusive RNA structures, one of

which forms a transcriptional terminator and results in premature termination,

and the other forms an antiterminator that allows read-through into the coding

sequence to produce a full-length mRNA. We developed a machine-learning based

approach, which, given a 5'UTR of a gene, predicts whether it can form the two

alternative structures typical to riboregulators employing conditional

termination. Using a large positive training set of riboregulators derived from

89 human microbiome bacteria, we show high specificity and sensitivity for our

classifier. We further show that our approach allows the discovery of previously

unidentified riboregulators, as exemplified by the detection of new LeuA leaders

and T-boxes in Streptococci Finally, we developed PASIFIC

(www.weizmann.ac.il/molgen/Sorek/PASIFIC/), an online web-server that, given a

user-provided 5'UTR sequence, predicts whether this sequence can adopt two

alternative structures conforming with the conditional termination paradigm. This

webserver is expected to assist in the identification of new riboswitches and

attenuators in the bacterial pan-genome.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw749

PMCID: PMC5314783

PMID: 27574119

11. Bioinformatics. 2017 Jan 15;33(2):266-271. doi: 10.1093/bioinformatics/btw612.

Epub 2016 Sep 25.

cMapper: gene-centric connectivity mapper for EBI-RDF platform.

Shoaib M(1,)(2), Ansari AA(1,)(2), Ahn SM(2,)(3,)(4).

Author information:

(1)Department of Biomedical Engineering, College of Medicine, University of

Ulsan, Asan Medical Center, Seoul, Republic of Korea. (2)Gachon Institute of

Genome Medicine and Sciences. (3)Division of Hematology and Oncology, Department

of Internal Medicine, Gachon University Gil Medical Center, Incheon, Republic of

Korea. (4)Department of Genome Medicine and Science, College of Medicine, Gachon

University, Seongnam, Republic of Korea.

MOTIVATION: In this era of biological big data, data integration has become a

common task and a challenge for biologists. The Resource Description Framework

(RDF) was developed to enable interoperability of heterogeneous datasets. The

EBI-RDF platform enables an efficient data integration of six independent

biological databases using RDF technologies and shared ontologies. However, to

take advantage of this platform, biologists need to be familiar with RDF

technologies and SPARQL query language. To overcome this practical limitation of

the EBI-RDF platform, we developed cMapper, a web-based tool that enables

biologists to search the EBI-RDF databases in a gene-centric manner without a

thorough knowledge of RDF and SPARQL.

RESULTS: cMapper allows biologists to search data entities in the EBI-RDF

platform that are connected to genes or small molecules of interest in multiple

biological contexts. The input to cMapper consists of a set of genes or small

molecules, and the output are data entities in six independent EBI-RDF databases

connected with the given genes or small molecules in the user's query. cMapper

provides output to users in the form of a graph in which nodes represent data

entities and the edges represent connections between data entities and inputted

set of genes or small molecules. Furthermore, users can apply filters based on

database, taxonomy, organ and pathways in order to focus on a core connectivity

graph of their interest. Data entities from multiple databases are differentiated

based on background colors. cMapper also enables users to investigate shared

connections between genes or small molecules of interest. Users can view the

output graph on a web browser or download it in either GraphML or JSON formats.

AVAILABILITY AND IMPLEMENTATION: cMapper is available as a web application with

an integrated MySQL database. The web application was developed using Java and

deployed on Tomcat server. We developed the user interface using HTML5, JQuery

and the Cytoscape Graph API. cMapper can be accessed at

http://cmapper.ewostech.net Readers can download the development manual from the

website http://cmapper.ewostech.net/docs/cMapperDocumentation.pdf. Source Code is

available at

https://github.com/muhammadshoaib/cmapperContact:smahn@gachon.ac.krSupplementary

information: Supplementary data are available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw612

PMID: 27667790

12. Bioinformatics. 2017 Jan 15;33(2):286-288. doi: 10.1093/bioinformatics/btw561.

Epub 2016 Aug 24.

Cas-analyzer: an online tool for assessing genome editing results using NGS data.

Park J(1), Lim K(2,)(3), Kim JS(2,)(3), Bae S(1,)(4).

Author information:

(1)Department of Chemistry, Hanyang University, Seoul, South Korea. (2)Center for

Genome Engineering, Institute for Basic Science, Seoul, South Korea.

(3)Department of Chemistry, Seoul National University, Seoul, South Korea and.

(4)Research Institute for Convergence of Basic Sciences, Hanyang University,

Seoul, South Korea.

Genome editing with programmable nucleases has been widely adopted in research

and medicine. Next generation sequencing (NGS) platforms are now widely used for

measuring the frequencies of mutations induced by CRISPR-Cas9 and other

programmable nucleases. Here, we present an online tool, Cas-Analyzer, a

JavaScript-based implementation for NGS data analysis. Because Cas-Analyzer is

completely used at a client-side web browser on-the-fly, there is no need to

upload very large NGS datasets to a server, a time-consuming step in genome

editing analysis. Currently, Cas-Analyzer supports various programmable

nucleases, including single nucleases and paired nucleases.AVAILABILITY AND

IMPLEMENTATION: Free access at http://www.rgenome.net/cas-analyzer/ CONTACT:

sangsubae@hanyang.ac.kr or jskim01@snu.ac.krSupplementary information:

Supplementary data are available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw561

PMCID: PMC5254075

PMID: 27559154

13. Nucleic Acids Res. 2017 Jan 9;45(1):e5. doi: 10.1093/nar/gkw819. Epub 2016 Sep

14.

Topology independent comparison of RNA 3D structures using the CLICK algorithm.

Nguyen MN(1), Sim AY(2), Wan Y(3), Madhusudhan MS(2,)(4), Verma C(2,)(5,)(6).

Author information:

(1)Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore 138671

minhn@bii.a-star.edu.sg. (2)Bioinformatics Institute, 30 Biopolis Street, #07-01,

Matrix, Singapore 138671. (3)Genome Institute of Singapore, 60 Biopolis Street,

Genome, #02-01, Singapore 138672. (4)Indian Institute of Science Education and

Research, Pune, India. (5)Department of Biological Sciences, National University

of Singapore, Singapore. (6)School of Biological Sciences, Nanyang Technological

University, Singapore.

RNA molecules are attractive therapeutic targets because non-coding RNA molecules

have increasingly been found to play key regulatory roles in the cell. Comparing

and classifying RNA 3D structures yields unique insights into RNA evolution and

function. With the rapid increase in the number of atomic-resolution RNA

structures, it is crucial to have effective tools to classify RNA structures and

to investigate them for structural similarities at different resolutions. We

previously developed the algorithm CLICK to superimpose a pair of protein 3D

structures by clique matching and 3D least squares fitting. In this study, we

extend and optimize the CLICK algorithm to superimpose pairs of RNA 3D structures

and RNA-protein complexes, independent of the associated topologies. Benchmarking

Rclick on four different datasets showed that it is either comparable to or

better than other structural alignment methods in terms of the extent of

structural overlaps. Rclick also recognizes conformational changes between RNA

structures and produces complementary alignments to maximize the extent of

detectable similarity. Applying Rclick to study Ribonuclease III protein

correctly aligned the RNA binding sites of RNAse III with its substrate. Rclick

can be further extended to identify ligand-binding pockets in RNA. A web server

is developed at http://mspc.bii.a-star.edu.sg/minhn/rclick.html.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw819

PMID: 27634929

14. Nucleic Acids Res. 2017 Jan 4;45(D1):D1040-D1045. doi: 10.1093/nar/gkw982. Epub

2016 Oct 24.

PlantTFDB 4.0: toward a central hub for transcription factors and regulatory

interactions in plants.

Jin J(1), Tian F(1,)(2), Yang DC(1), Meng YQ(1), Kong L(1), Luo J(3), Gao G(4).

Author information:

(1)State Key Laboratory of Protein and Plant Gene Research, College of Life

Sciences, Center for Bioinformatics, Beijing 100871, P.R. China.

(2)Peking-Tsinghua Center for Life Sciences, Academy for Advanced

Interdisciplinary Studies, Peking University, Beijing 100871, P.R. China.

(3)State Key Laboratory of Protein and Plant Gene Research, College of Life

Sciences, Center for Bioinformatics, Beijing 100871, P.R. China

luojc@mail.cbi.pku.edu.cn. (4)State Key Laboratory of Protein and Plant Gene

Research, College of Life Sciences, Center for Bioinformatics, Beijing 100871,

P.R. China gaog@mail.cbi.pku.edu.cn.

With the goal of providing a comprehensive, high-quality resource for both plant

transcription factors (TFs) and their regulatory interactions with target genes,

we upgraded plant TF database PlantTFDB to version 4.0

(http://planttfdb.cbi.pku.edu.cn/). In the new version, we identified 320 370 TFs

from 165 species, presenting a more comprehensive genomic TF repertoires of green

plants. Besides updating the pre-existing abundant functional and evolutionary

annotation for identified TFs, we generated three new types of annotation which

provide more directly clues to investigate functional mechanisms underlying: (i)

a set of high-quality, non-redundant TF binding motifs derived from experiments;

(ii) multiple types of regulatory elements identified from high-throughput

sequencing data; (iii) regulatory interactions curated from literature and

inferred by combining TF binding motifs and regulatory elements. In addition, we

upgraded previous TF prediction server, and set up four novel tools for

regulation prediction and functional enrichment analyses. Finally, we set up a

novel companion portal PlantRegMap (http://plantregmap.cbi.pku.edu.cn) for users

to access the regulation resource and analysis tools conveniently.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw982

PMCID: PMC5210657

PMID: 27924042

15. Nucleic Acids Res. 2017 Jan 4;45(D1):D68-D73. doi: 10.1093/nar/gkw925. Epub 2016

Oct 18.

L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes.

Penzkofer T(1), Jäger M(2), Figlerowicz M(3), Badge R(4), Mundlos S(2), Robinson

PN(2,)(5), Zemojtel T(6,)(3).

Author information:

(1)Department of Radiology, Charité-Universitätsmedizin Berlin, Augustenburger

Platz 1, 13353 Berlin, Germany. (2)Institut für Medizinische Genetik und

Humangenetik, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353

Berlin, Germany. (3)Institute of Bioorganic Chemistry, Polish Academy of

Sciences, 60-569 Poznan, Poland. (4)Department of Genetics, University of

Leicester, Leicester, LE1 7RH, UK. (5)The Jackson Laboratory for Genomic

medicine, 10 Discovery Drive, Farmington, CT 06032, USA. (6)Institut für

Medizinische Genetik und Humangenetik, Charité-Universitätsmedizin Berlin,

Augustenburger Platz 1, 13353 Berlin, Germany tomasz.zemojtel@charite.de.

LINE-1 (L1) insertions comprise as much as 17% of the human genome sequence, and

similar proportions have been recorded for other mammalian species. Given the

established role of L1 retrotransposons in shaping mammalian genomes, it becomes

an important task to track and annotate the sources of this activity: full length

elements, able to encode the cis and trans acting components of the

retrotransposition machinery. The L1Base database (http://l1base.charite.de)

contains annotated full-length sequences of LINE-1 transposons including

putatively active L1s. For the new version of L1Base, a LINE-1 annotation tool,

L1Xplorer, has been used to mine potentially active L1 retrotransposons from the

reference genome sequences of 17 mammals. The current release of the human

genome, GRCh38, contains 146 putatively active L1 elements or full length intact

L1 elements (FLIs). The newest versions of the mouse, GRCm38 and the rat,

Rnor\_6.0, genomes contain 2811 and 492 FLIs, respectively. Most likely reflecting

the current level of completeness of the genome project, the latest reference

sequence of the common chimpanzee genome, PT 2.19, only contains 19 FLIs. Of

note, the current assemblies of the dog, CF 3.1 and the sheep, OA 3.1, genomes

contain 264 and 598 FLIs, respectively. Further developments in the new version

of L1Base include an updated website with implementation of modern web server

technologies. including a more responsive design for an improved user experience,

as well as the addition of data sharing capabilities for L1Xplorer annotation.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw925

PMCID: PMC5210629

PMID: 27924012

16. Nucleic Acids Res. 2017 Jan 4;45(D1):D658-D662. doi: 10.1093/nar/gkw983. Epub

2016 Oct 26.

Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility

data in human and mouse.

Mei S(1,)(2), Qin Q(1,)(2), Wu Q(1,)(2), Sun H(2), Zheng R(2), Zang C(3,)(4), Zhu

M(2), Wu J(5), Shi X(2), Taing L(3), Liu T(6), Brown M(4,)(7), Meyer CA(8,)(4),

Liu XS(9,)(2,)(3,)(4).

Author information:

(1)Clinical Translational Research Center, Shanghai Pulmonary Hospital, Tongji

University, Shanghai 200433, China. (2)Department of Bioinformatics, School of

Life Sciences and Technology, Tongji University, Shanghai 200092, China.

(3)Department of Biostatistics and Computational Biology, Dana-Farber Cancer

Institute and Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA.

(4)Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute,

Boston, MA 02215, USA. (5)MOE Key Laboratory of Bioinformatics, Bioinformatics

Division and Center for Synthetic & Systems Biology, TNLIST; Department of

Automation, Tsinghua University, Beijing 100084, China. (6)Department of

Biochemistry, University at Buffalo, Buffalo, NY 14214, USA. (7)Department of

Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School,

Boston, MA 02215, USA. (8)Department of Biostatistics and Computational Biology,

Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health,

Boston, MA 02215, USA cliff@jimmy.harvard.edu. (9)Clinical Translational Research

Center, Shanghai Pulmonary Hospital, Tongji University, Shanghai 200433, China

xsliu@jimmy.harvard.edu.

Chromatin immunoprecipitation, DNase I hypersensitivity and

transposase-accessibility assays combined with high-throughput sequencing enable

the genome-wide study of chromatin dynamics, transcription factor binding and

gene regulation. Although rapidly accumulating publicly available ChIP-seq,

DNase-seq and ATAC-seq data are a valuable resource for the systematic

investigation of gene regulation processes, a lack of standardized curation,

quality control and analysis procedures have hindered extensive reuse of these

data. To overcome this challenge, we built the Cistrome database, a collection of

ChIP-seq and chromatin accessibility data (DNase-seq and ATAC-seq) published

before January 1, 2016, including 13 366 human and 9953 mouse samples. All the

data have been carefully curated and processed with a streamlined analysis

pipeline and evaluated with comprehensive quality control metrics. We have also

created a user-friendly web server for data query, exploration and visualization.

The resulting Cistrome DB (Cistrome Data Browser), available online at

http://cistrome.org/db, is expected to become a valuable resource for

transcriptional and epigenetic regulation studies.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw983

PMCID: PMC5210658

PMID: 27789702

17. Nucleic Acids Res. 2017 Jan 4;45(D1):1015-1020. doi: 10.1093/nar/gkw935. Epub

2016 Oct 13.

PIECE 2.0: an update for the plant gene structure comparison and evolution

database.

Wang Y(1,)(2,)(3), Xu L(2,)(3), Thilmony R(1), You FM(4), Gu YQ(5), Coleman-Derr

D(6,)(3).

Author information:

(1)USDA-ARS, Western Regional Research Center, Crop Improvement and Genetics

Research Unit, Albany, CA 94710, USA. (2)Department of Plant and Microbial

Biology, University of California Berkeley, Berkeley, CA, USA. (3)USDA-ARS, Plant

Gene Expression Center, Albany, CA 94710, USA. (4)Cereal Research Centre,

Agriculture and Agri-Food Canada, Morden R6M 1Y5 MB, Canada. (5)USDA-ARS, Western

Regional Research Center, Crop Improvement and Genetics Research Unit, Albany, CA

94710, USA yong.gu@ars.usda.gov. (6)Department of Plant and Microbial Biology,

University of California Berkeley, Berkeley, CA, USA

devin.coleman-derr@ars.usda.gov.

PIECE (Plant Intron Exon Comparison and Evolution) is a web-accessible database

that houses intron and exon information of plant genes. PIECE serves as a

resource for biologists interested in comparing intron-exon organization and

provides valuable insights into the evolution of gene structure in plant genomes.

Recently, we updated PIECE to a new version, PIECE 2.0

(http://probes.pw.usda.gov/piece or http://aegilops.wheat.ucdavis.edu/piece).

PIECE 2.0 contains annotated genes from 49 sequenced plant species as compared to

25 species in the previous version. In the current version, we also added several

new features: (i) a new viewer was developed to show phylogenetic trees displayed

along with the structure of individual genes; (ii) genes in the phylogenetic tree

can now be also grouped according to KOG (The annotation of Eukaryotic

Orthologous Groups) and KO (KEGG Orthology) in addition to Pfam domains; (iii)

information on intronless genes are now included in the database; (iv) a

statistical summary of global gene structure information for each species and its

comparison with other species was added; and (v) an improved GSDraw tool was

implemented in the web server to enhance the analysis and display of gene

structure. The updated PIECE 2.0 database will be a valuable resource for the

plant research community for the study of gene structure and evolution.

© Crown copyright 2016.

DOI: 10.1093/nar/gkw935

PMCID: PMC5210635

PMID: 27742820

18. Nucleic Acids Res. 2017 Jan 4;45(D1):D347-D352. doi: 10.1093/nar/gkw918. Epub

2016 Oct 12.

Ontobee: A linked ontology data server to support ontology term dereferencing,

linkage, query and integration.

Ong E(1), Xiang Z(1), Zhao B(1), Liu Y(1), Lin Y(1), Zheng J(2), Mungall C(3),

Courtot M(4), Ruttenberg A(5), He Y(6).

Author information:

(1)University of Michigan Medical School, Ann Arbor, MI 48109, USA. (2)University

of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.

(3)Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. (4)European

Molecular Biology Laboratory-European Bioinformatics Institute, Hinxton,

Cambridge, CB10 1SD, UK. (5)University at Buffalo, Buffalo, NY 14203, USA.

(6)University of Michigan Medical School, Ann Arbor, MI 48109, USA

yongqunh@umich.edu.

Linked Data (LD) aims to achieve interconnected data by representing entities

using Unified Resource Identifiers (URIs), and sharing information using Resource

Description Frameworks (RDFs) and HTTP. Ontologies, which logically represent

entities and relations in specific domains, are the basis of LD. Ontobee

(http://www.ontobee.org/) is a linked ontology data server that stores ontology

information using RDF triple store technology and supports query, visualization

and linkage of ontology terms. Ontobee is also the default linked data server for

publishing and browsing biomedical ontologies in the Open Biological Ontology

(OBO) Foundry (http://obofoundry.org) library. Ontobee currently hosts more than

180 ontologies (including 131 OBO Foundry Library ontologies) with over four

million terms. Ontobee provides a user-friendly web interface for querying and

visualizing the details and hierarchy of a specific ontology term. Using the

eXtensible Stylesheet Language Transformation (XSLT) technology, Ontobee is able

to dereference a single ontology term URI, and then output RDF/eXtensible Markup

Language (XML) for computer processing or display the HTML information on a web

browser for human users. Statistics and detailed information are generated and

displayed for each ontology listed in Ontobee. In addition, a SPARQL web

interface is provided for custom advanced SPARQL queries of one or multiple

ontologies.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw918

PMCID: PMC5210626

PMID: 27733503

19. Nucleic Acids Res. 2017 Jan 4;45(D1):D1082-D1089. doi: 10.1093/nar/gkw704. Epub

2016 Aug 4.

SoyNet: a database of co-functional networks for soybean Glycine max.

Kim E(1), Hwang S(1), Lee I(2).

Author information:

(1)Department of Biotechnology, College of Life Science and Biotechnology, Yonsei

University, Seoul, Korea. (2)Department of Biotechnology, College of Life Science

and Biotechnology, Yonsei University, Seoul, Korea insuklee@yonsei.ac.kr.

Soybean (Glycine max) is a legume crop with substantial economic value, providing

a source of oil and protein for humans and livestock. More than 50% of edible

oils consumed globally are derived from this crop. Soybean plants are also

important for soil fertility, as they fix atmospheric nitrogen by symbiosis with

microorganisms. The latest soybean genome annotation (version 2.0) lists 56 044

coding genes, yet their functional contributions to crop traits remain mostly

unknown. Co-functional networks have proven useful for identifying genes that are

involved in a particular pathway or phenotype with various network algorithms.

Here, we present SoyNet (available at www.inetbio.org/soynet), a database of

co-functional networks for G. max and a companion web server for network-based

functional predictions. SoyNet maps 1 940 284 co-functional links between 40 812

soybean genes (72.8% of the coding genome), which were inferred from 21 distinct

types of genomics data including 734 microarrays and 290 RNA-seq samples from

soybean. SoyNet provides a new route to functional investigation of the soybean

genome, elucidating genes and pathways of agricultural importance.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw704

PMCID: PMC5210602

PMID: 27492285

20. Bioinformatics. 2017 Jan 1;33(1):148-149. doi: 10.1093/bioinformatics/btw579.

Epub 2016 Sep 6.

OntoBrowser: a collaborative tool for curation of ontologies by subject matter

experts.

Ravagli C(1), Pognan F(1), Marc P(1).

Author information:

(1)PreClinical Safety, Translational Sciences, Novartis Institute for Biomedical

Research, Basel, CH-4002, Switzerland.

The lack of controlled terminology and ontology usage leads to incomplete search

results and poor interoperability between databases. One of the major underlying

challenges of data integration is curating data to adhere to controlled

terminologies and/or ontologies. Finding subject matter experts with the time and

skills required to perform data curation is often problematic. In addition,

existing tools are not designed for continuous data integration and collaborative

curation. This results in time-consuming curation workflows that often become

unsustainable. The primary objective of OntoBrowser is to provide an easy-to-use

online collaborative solution for subject matter experts to map reported terms to

preferred ontology (or code list) terms and facilitate ontology evolution.

Additional features include web service access to data, visualization of

ontologies in hierarchical/graph format and a peer review/approval workflow with

alerting.AVAILABILITY AND IMPLEMENTATION: The source code is freely available

under the Apache v2.0 license. Source code and installation instructions are

available at http://opensource.nibr.com This software is designed to run on a

Java EE application server and store data in a relational database.

CONTACT: philippe.marc@novartis.com.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw579

PMID: 27605099

21. Bioinformatics. 2017 Jan 1;33(1):122-124. doi: 10.1093/bioinformatics/btw564.

Epub 2016 Aug 26.

PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids

composition.

Zuo Y(1), Li Y(1,)(2), Chen Y(3), Li G(1), Yan Z(1,)(3), Yang L(4).

Author information:

(1)The Key Laboratory of Mammalian Reproductive Biology and Biotechnology of the

Ministry of Education, College of Life Sciences, Inner Mongolia University,

Hohhot 010021, China. (2)Department of Mechanical Engineering, Columbia

University, New York, NY 10027, USA. (3)School of Physical Science and

Technology, Inner Mongolia University, Hohhot 010021, China. (4)College of

Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081,

China.

The reduced amino acids perform powerful ability for both simplifying protein

complexity and identifying functional conserved regions. However, dealing with

different protein problems may need different kinds of cluster methods.

Encouraged by the success of pseudo-amino acid composition algorithm, we

developed a freely available web server, called PseKRAAC (the pseudo K-tuple

reduced amino acids composition). By implementing reduced amino acid alphabets,

the protein complexity can be significantly simplified, which leads to decrease

chance of overfitting, lower computational handicap and reduce information

redundancy. PseKRAAC delivers more capability for protein research by

incorporating three crucial parameters that describes protein composition. Users

can easily generate many different modes of PseKRAAC tailored to their needs by

selecting various reduced amino acids alphabets and other characteristic

parameters. It is anticipated that the PseKRAAC web server will become a very

useful tool in computational proteomics and protein sequence

analysis.AVAILABILITY AND IMPLEMENTATION: Freely available on the web at

http://bigdata.imu.edu.cn/psekraac CONTACTS: yczuo@imu.edu.cn or

imu.hema@foxmail.com or yanglei\_hmu@163.comSupplementary information:

Supplementary data are available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw564

PMID: 27565583

22. Bioinformatics. 2017 Jan 1;33(1):35-41. doi: 10.1093/bioinformatics/btw539. Epub

2016 Aug 16.

iRSpot-EL: identify recombination spots with an ensemble learning approach.

Liu B(1,)(2,)(3), Wang S(1), Long R(1), Chou KC(3,)(4).

Author information:

(1)School of Computer Science and Technology. (2)Key Laboratory of Network

Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, Guangdong 518055, China. (3)Gordon Life Science

Institute, Belmont, MA 02478, USA. (4)Center of Excellence in Genomic Medicine

Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia.

MOTIVATION: Coexisting in a DNA system, meiosis and recombination are two

indispensible aspects for cell reproduction and growth. With the avalanche of

genome sequences emerging in the post-genomic age, it is an urgent challenge to

acquire the information of DNA recombination spots because it can timely provide

very useful insights into the mechanism of meiotic recombination and the process

of genome evolution.

RESULTS: To address such a challenge, we have developed a predictor, called

IRSPOT-EL: , by fusing different modes of pseudo K-tuple nucleotide composition

and mode of dinucleotide-based auto-cross covariance into an ensemble classifier

of clustering approach. Five-fold cross tests on a widely used benchmark dataset

have indicated that the new predictor remarkably outperforms its existing

counterparts. Particularly, far beyond their reach, the new predictor can be

easily used to conduct the genome-wide analysis and the results obtained are

quite consistent with the experimental map.

AVAILABILITY AND IMPLEMENTATION: For the convenience of most experimental

scientists, a user-friendly web-server for iRSpot-EL has been established at

http://bioinformatics.hitsz.edu.cn/iRSpot-EL/, by which users can easily obtain

their desired results without the need to go through the complicated mathematical

equations involved.

CONTACT: bliu@gordonlifescience.org or bliu@insun.hit.edu.cnSupplementary

information: Supplementary data are available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw539

PMID: 27531102

23. Chem Biol Drug Des. 2017 Jan;89(1):74-83. doi: 10.1111/cbdd.12834. Epub 2016 Sep

9.

AVCpred: an integrated web server for prediction and design of antiviral

compounds.

Qureshi A(1), Kaur G(1), Kumar M(1).

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Council of

Scientific and Industrial Research, Chandigarh, India.

Viral infections constantly jeopardize the global public health due to lack of

effective antiviral therapeutics. Therefore, there is an imperative need to speed

up the drug discovery process to identify novel and efficient drug candidates. In

this study, we have developed quantitative structure-activity relationship

(QSAR)-based models for predicting antiviral compounds (AVCs) against deadly

viruses like human immunodeficiency virus (HIV), hepatitis C virus (HCV),

hepatitis B virus (HBV), human herpesvirus (HHV) and 26 others using publicly

available experimental data from the ChEMBL bioactivity database. Support vector

machine (SVM) models achieved a maximum Pearson correlation coefficient of 0.72,

0.74, 0.66, 0.68, and 0.71 in regression mode and a maximum Matthew's correlation

coefficient 0.91, 0.93, 0.70, 0.89, and 0.71, respectively, in classification

mode during 10-fold cross-validation. Furthermore, similar performance was

observed on the independent validation sets. We have integrated these models in

the AVCpred web server, freely available at http://crdd.osdd.net/servers/avcpred.

In addition, the datasets are provided in a searchable format. We hope this web

server will assist researchers in the identification of potential antiviral

agents. It would also save time and cost by prioritizing new drugs against

viruses before their synthesis and experimental testing.

© 2016 John Wiley & Sons A/S.

DOI: 10.1111/cbdd.12834

PMID: 27490990

24. Hum Mutat. 2017 Jan;38(1):25-33. doi: 10.1002/humu.23125. Epub 2016 Oct 13.

mirVAFC: A Web Server for Prioritizations of Pathogenic Sequence Variants from

Exome Sequencing Data via Classifications.

Li Z(1), Liu Z(1), Jiang Y(1), Chen D(1), Ran X(1), Sun ZS(1,)(2), Wu J(1).

Author information:

(1)Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, China.

(2)Beijing Institute of Life Science, Chinese Academy of Sciences, Beijing,

China.

Exome sequencing has been widely used to identify the genetic variants underlying

human genetic disorders for clinical diagnoses, but the identification of

pathogenic sequence variants among the huge amounts of benign ones is complicated

and challenging. Here, we describe a new Web server named mirVAFC for pathogenic

sequence variants prioritizations from clinical exome sequencing (CES) variant

data of single individual or family. The mirVAFC is able to comprehensively

annotate sequence variants, filter out most irrelevant variants using custom

criteria, classify variants into different categories as for estimated

pathogenicity, and lastly provide pathogenic variants prioritizations based on

classifications and mutation effects. Case studies using different types of

datasets for different diseases from publication and our in-house data have

revealed that mirVAFC can efficiently identify the right pathogenic candidates as

in original work in each case. Overall, the Web server mirVAFC is specifically

developed for pathogenic sequence variant identifications from family-based CES

variants using classification-based prioritizations. The mirVAFC Web server is

freely accessible at https://www.wzgenomics.cn/mirVAFC/.

© 2016 WILEY PERIODICALS, INC.

DOI: 10.1002/humu.23125

PMID: 27676360

25. Infect Genet Evol. 2017 Jan;47:132-139. doi: 10.1016/j.meegid.2016.10.008. Epub

2016 Oct 17.

The wing venation patterns to identify single tsetse flies.

Kaba D(1), Berté D(1), Ta BT(1), Tellería J(2), Solano P(2), Dujardin JP(2).

Author information:

(1)INSP/IPR Bouaké, Côte d'Ivoire. (2)UMR 177, INTERTRYP, CIRAD-IRD, Baillarguet,

France.

This is the first study to explore the potential of various geometric

morphometrics methods to help the morphological diagnostic of tsetse species,

vectors of human and animal trypanosomiases in sub-Saharan Africa. We compared

landmarks, semilandmarks and outlines techniques on male and female samples of

species, and suggested adapted strategies according to the countries and their

own Glossina fauna. We could compare up to 7 taxa belonging to the three main

subgenera of the Glossina genus: Nemorhina (5 species), Glossina (1 species) and

Austenina (1 species). Our sample included the major vectors of sleeping

sickness: G. palpalis palpalis, G. p. gambiensis, G. fuscipes fuscipes and G. f.

quanzensis, as well as two important vectors of African animal trypanosomoses: G.

tachinoides and Glossina morsitans submorsitans. The average level of correct

species recognition by the wing shape was satisfactory, and slightly higher for

females than for males. The best scores of correct assignment, in both sexes,

were obtained by the contour technique (96% of correct attribution in females,

92% in males), slightly higher than for semilandmarks (95% and 91%) or landmarks

(94% and 89%) techniques. We made our images of wings freely available to be used

as reference images (http://mome-clic.com), and we describe the conditions and

the analytical steps to be followed to identify unknown specimens using external

reference images. Under adequate conditions, such use of reference images

obtained from a free access server could help species identification of new

samples anywhere in Africa.

Copyright © 2016 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.meegid.2016.10.008

PMID: 27765637

26. Proteins. 2017 Jan;85(1):72-77. doi: 10.1002/prot.25199. Epub 2016 Nov 13.

Aquerium: A web application for comparative exploration of domain-based protein

occurrences on the taxonomically clustered genome tree.

Adebali O(1,)(2,)(3), Zhulin IB(1,)(2,)(3).

Author information:

(1)UT-ORNL Graduate School of Genome Science and Technology, University of

Tennessee, Knoxville, Tennessee, 37996. (2)Department of Microbiology, University

of Tennessee, Knoxville, Tennessee, 37996. (3)Computer Science and Mathematics

Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37961.

Gene duplication and loss are major driving forces in evolution. While many

important genomic resources provide information on gene presence, there is a lack

of tools giving equal importance to presence and absence information as well as

web platforms enabling easy visual comparison of multiple domain-based protein

occurrences at once. Here, we present Aquerium, a platform for visualizing

genomic presence and absence of biomolecules with a focus on protein domain

architectures. The web server offers advanced domain organization querying

against the database of pre-computed domains for ∼26,000 organisms and it can be

utilized for identification of evolutionary events, such as fusion,

disassociation, duplication, and shuffling of protein domains. The tool also

allows alternative inputs of custom entries or BLASTP results for visualization.

Aquerium will be a useful tool for biologists who perform comparative genomic and

evolutionary analyses. The web server is freely accessible at

http://aquerium.utk.edu. Proteins 2016; 85:72-77. © 2016 Wiley Periodicals, Inc.

© 2016 Wiley Periodicals, Inc.

DOI: 10.1002/prot.25199

PMCID: PMC5167639 [Available on 2018-01-01]

PMID: 27802571

27. RNA. 2017 Jan;23(1):14-22. Epub 2016 Nov 2.

Genome-scale characterization of RNA tertiary structures and their functional

impact by RNA solvent accessibility prediction.

Yang Y(1), Li X(1,)(2), Zhao H(3), Zhan J(1), Wang J(4), Zhou Y(1,)(4).

Author information:

(1)Institute for Glycomics and School of Information and Communication

Technology, Griffith University, Gold Coast, QLD 4222, Australia. (2)School of

Computer Science and Information Engineering, Hefei University of Technology,

Hefei 230009, China. (3)Institute of Health and Biomedical Innovation, Queensland

University of Technology, Queensland 4222, Australia. (4)Shandong Provincial Key

Laboratory of Biophysics, Institute of Biophysics, Dezhou University, Dezhou

253023, China.

As most RNA structures are elusive to structure determination, obtaining solvent

accessible surface areas (ASAs) of nucleotides in an RNA structure is an

important first step to characterize potential functional sites and core

structural regions. Here, we developed RNAsnap, the first machine-learning method

trained on protein-bound RNA structures for solvent accessibility prediction.

Built on sequence profiles from multiple sequence alignment (RNAsnap-prof), the

method provided robust prediction in fivefold cross-validation and an independent

test (Pearson correlation coefficients, r, between predicted and actual ASA

values are 0.66 and 0.63, respectively). Application of the method to 6178 mRNAs

revealed its positive correlation to mRNA accessibility by dimethyl sulphate

(DMS) experimentally measured in vivo (r = 0.37) but not in vitro (r = 0.07),

despite the lack of training on mRNAs and the fact that DMS accessibility is only

an approximation to solvent accessibility. We further found strong association

across coding and noncoding regions between predicted solvent accessibility of

the mutation site of a single nucleotide variant (SNV) and the frequency of that

variant in the population for 2.2 million SNVs obtained in the 1000 Genomes

Project. Moreover, mapping solvent accessibility of RNAs to the human genome

indicated that introns, 5' cap of 5' and 3' cap of 3' untranslated regions, are

more solvent accessible, consistent with their respective functional roles. These

results support conformational selections as the mechanism for the formation of

RNA-protein complexes and highlight the utility of genome-scale characterization

of RNA tertiary structures by RNAsnap. The server and its stand-alone

downloadable version are available at http://sparks-lab.org.

© 2016 Yang et al.; Published by Cold Spring Harbor Laboratory Press for the RNA

Society.

DOI: 10.1261/rna.057364.116

PMCID: PMC5159645

PMID: 27807179

28. Bioinformatics. 2016 Dec 28. pii: btw694. doi: 10.1093/bioinformatics/btw694.

[Epub ahead of print]

QAcon: single model quality assessment using protein structural and contact

information with machine learning techniques.

Cao R(1), Adhikari B(2), Bhattacharya D(3), Sun M(4), Hou J(2), Cheng J(2,)(5).

Author information:

(1)Department of Computer Science, Pacific Lutheran University, WA 98447, USA.

(2)Department of Computer Science, University of Missouri, Columbia, MO 65211,

USA. (3)Department of Electrical Engineering and Computer Science, Wichita State

University, Wichita, KS 67260-0083, USA. (4)Department of Electrical and Computer

Engineering. (5)Informatics Institute, University of Missouri, Columbia, MO

65211, USA.

MOTIVATION: Protein model quality assessment (QA) plays a very important role in

protein structure prediction. It can be divided into two groups of methods:

single model and consensus QA method. The consensus QA methods may fail when

there is a large portion of low quality models in the model pool.

RESULTS: In this paper, we develop a novel single-model quality assessment method

QAcon utilizing structural features, physicochemical properties, and residue

contact predictions. We apply residue-residue contact information predicted by

two protein contact prediction methods PSICOV and DNcon to generate a new score

as feature for quality assessment. This novel feature and other 11 features are

used as input to train a two-layer neural network on CASP9 datasets to predict

the quality of a single protein model. We blindly benchmarked our method QAcon on

CASP11 dataset as the MULTICOM-CLUSTER server. Based on the evaluation, our

method is ranked as one of the top single model QA methods. The good performance

of the features based on contact prediction illustrates the value of using

contact information in protein quality assessment.

AVAILABILITY AND IMPLEMENTATION: The web server and the source code of QAcon are

freely available at: http://cactus.rnet.missouri.edu/QAcon CONTACT:

chengji@missouri.eduSupplementary information: Supplementary data are available

at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw694

PMID: 28035027

29. Bioinformatics. 2016 Dec 28. pii: btw790. doi: 10.1093/bioinformatics/btw790.

[Epub ahead of print]

dAPE: a web server to detect homorepeats and follow their evolution.

Mier P(1), Andrade-Navarro MA(1).

Author information:

(1)Faculty of Biology, Johannes Gutenberg Universität, Institute of Molecular

Biology, Mainz 55128, Germany.

Homorepeats are low complexity regions consisting of repetitions of a single

amino acid residue. There is no current consensus on the minimum number of

residues needed to define a functional homorepeat, nor even if mismatches are

allowed. Here we present dAPE, a web server that helps following the evolution of

homorepeats based on orthology information, using a sensitive but tunable cutoff

to help in the identification of emerging homorepeats.AVAILABILITY AND

IMPLEMENTATION: dAPE can be accessed from

http://cbdm-01.zdv.uni-mainz.de/∼munoz/polyx CONTACT:

munoz@uni-mainz.deSupplementary information: Supplementary data are available at

Bioinformatics online.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw790

PMID: 28031183

30. J Chem Inf Model. 2016 Dec 27;56(12):2287-2291. doi: 10.1021/acs.jcim.6b00407.

Epub 2016 Dec 15.

MIB: Metal Ion-Binding Site Prediction and Docking Server.

Lin YF(1), Cheng CW(1), Shih CS(1), Hwang JK(1), Yu CS, Lu CH(2).

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University , Hsinchu, 30050 Taiwan. (2)Graduate Institute of Basic Medical

Science, China Medical University , Taichung 40402, Taiwan.

The structure of a protein determines its biological function(s) and its

interactions with other factors; the binding regions tend to be conserved in

sequence and structure, and the interacting residues involved are usually in

close 3D space. The Protein Data Bank currently contains more than 110 000

protein structures, approximately one-third of which contain metal ions.

Identifying and characterizing metal ion-binding sites is thus essential for

investigating a protein's function(s) and interactions. However, experimental

approaches are time-consuming and costly. The web server reported here was built

to predict metal ion-binding residues and to generate the predicted metal

ion-bound 3D structure. Binding templates have been constructed for regions that

bind 12 types of metal ion-binding residues have been used to construct binding

templates. The templates include residues within 3.5 Å of the metal ion, and the

fragment transformation method was used for structural comparison between query

proteins and templates without any data training. Through the adjustment of

scoring functions, which are based on the similarity of structure and binding

residues. Twelve kinds of metal ions (Ca(2+), Cu(2+), Fe(3+), Mg(2+), Mn(2+),

Zn(2+), Cd(2+), Fe(2+), Ni(2+), Hg(2+), Co(2+), and Cu(+)) binding residues

prediction are supported. MIB also provides the metal ions docking after

prediction. The MIB server is available at http://bioinfo.cmu.edu.tw/MIB/ .

DOI: 10.1021/acs.jcim.6b00407

PMID: 27976886

31. Bioinformatics. 2016 Dec 24. pii: btw725. doi: 10.1093/bioinformatics/btw725.

[Epub ahead of print]

Calypso: a user-friendly web-server for mining and visualizing

microbiome-environment interactions.

Zakrzewski M(1), Proietti C(1), Ellis JJ(2), Hasan S(1,)(2), Brion MJ(2), Berger

B(3), Krause L(1,)(2).

Author information:

(1)QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia.

(2)The University of Queensland Diamantina Institute, Brisbane, QLD 4102,

Australia. (3)Nestle Research Center, Vers-chez-les-Blanc, Lausanne, Switzerland.

Calypso is an easy-to-use online software suite that allows non-expert users to

mine, interpret and compare taxonomic information from metagenomic or 16S rDNA

datasets. Calypso has a focus on multivariate statistical approaches that can

identify complex environment-microbiome associations. The software enables

quantitative visualizations, statistical testing, multivariate analysis,

supervised learning, factor analysis, multivariable regression, network analysis

and diversity estimates. Comprehensive help pages, tutorials and videos are

provided via a wiki page.AVAILABILITY AND IMPLEMENTATION: The web-interface is

accessible via http://cgenome.net/calypso/ The software is programmed in Java,

PERL and R and the source code is available from Zenodo

(https://zenodo.org/record/50931). The software is freely available for

non-commercial users.

CONTACT: l.krause@uq.edu.auSupplementary information: Supplementary data are

available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw725

PMID: 28025202

32. Bioinformatics. 2016 Dec 23. pii: btw717. doi: 10.1093/bioinformatics/btw717.

[Epub ahead of print]

FUEL-mLoc: feature-unified prediction and explanation of multi-localization of

cellular proteins in multiple organisms.

Wan S(1), Mak MW(2), Kung SY(1).

Author information:

(1)Department of Electrical Engineering, Princeton University, NJ, USA.

(2)Department of Electronic and Information Engineering, The Hong Kong

Polytechnic University, Hong Kong, SAR, China.

Although many web-servers for predicting protein subcellular localization have

been developed, they often have the following drawbacks: (i) lack of

interpretability or interpreting results with heterogenous information which may

confuse users; (ii) ignoring multi-location proteins and (iii) only focusing on

specific organism. To tackle these problems, we present an interpretable and

efficient web-server, namely FUEL-mLoc, using F: eature- U: nified prediction and

E: xplanation of multi- L: ocalization of cellular proteins in multiple

organisms. Compared to conventional localization predictors, FUEL-mLoc has the

following advantages: (i) using unified features (i.e. essential GO terms) to

interpret why a prediction is made; (ii) being capable of predicting both single-

and multi-location proteins and (iii) being able to handle proteins of multiple

organisms, including Eukaryota, Homo sapiens, Viridiplantae, Gram-positive

Bacteria, Gram-negative Bacteria and Virus Experimental results demonstrate that

FUEL-mLoc outperforms state-of-the-art subcellular-localization

predictors.AVAILABILITY AND IMPLEMENTATION:

http://bioinfo.eie.polyu.edu.hk/FUEL-mLoc/ CONTACTS: shibiao.wan@princeton.edu or

enmwmak@polyu.edu.hkSupplementary information: Supplementary data are available

at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw717

PMID: 28011780

33. BMC Bioinformatics. 2016 Dec 23;17(Suppl 17):469. doi: 10.1186/s12859-016-1328-7.

Multi-CAR: a tool of contig scaffolding using multiple references.

Chen KT(1), Chen CJ(1), Shen HT(1), Liu CL(1), Huang SH(1), Lu CL(2).

Author information:

(1)Department of Computer Science, National Tsing Hua University, Hsinchu, 30013,

Taiwan. (2)Department of Computer Science, National Tsing Hua University,

Hsinchu, 30013, Taiwan. cllu@cs.nthu.edu.tw.

BACKGROUND: A draft genome assembled by current next-generation sequencing

techniques from short reads is just a collection of contigs, whose relative

positions and orientations along the genome being sequenced are unknown. To

further obtain its complete sequence, a contig scaffolding process is usually

applied to order and orient the contigs in the draft genome. Although several

single reference-based scaffolding tools have been proposed, they may produce

erroneous scaffolds if there are rearrangements between the target and reference

genomes or their phylogenetic relationship is distant. This may suggest that a

single reference genome may not be sufficient to produce correct scaffolds of a

draft genome.

RESULTS: In this study, we design a simple heuristic method to further revise our

single reference-based scaffolding tool CAR into a new one called Multi-CAR such

that it can utilize multiple complete genomes of related organisms as references

to more accurately order and orient the contigs of a draft genome. In practical

usage, our Multi-CAR does not require prior knowledge concerning phylogenetic

relationships among the draft and reference genomes and libraries of paired-end

reads. To validate Multi-CAR, we have tested it on a real dataset composed of

several prokaryotic genomes and also compared its accuracy performance with other

multiple reference-based scaffolding tools Ragout and MeDuSa. Our experimental

results have finally shown that Multi-CAR indeed outperforms Ragout and MeDuSa in

terms of sensitivity, precision, genome coverage, scaffold number and scaffold

N50 size.

CONCLUSIONS: Multi-CAR serves as an efficient tool that can more accurately order

and orient the contigs of a draft genome based on multiple reference genomes. The

web server of Multi-CAR is freely available at

http://genome.cs.nthu.edu.tw/Multi-CAR/ .

DOI: 10.1186/s12859-016-1328-7

PMCID: PMC5260120

PMID: 28155633

34. BMC Bioinformatics. 2016 Dec 23;17(Suppl 17):468. doi: 10.1186/s12859-016-1325-x.

Global inference of disease-causing single nucleotide variants from exome

sequencing data.

Wu M(1,)(2), Chen T(1,)(2), Jiang R(3,)(4).

Author information:

(1)MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for

Synthetic & Systems Biology, Tsinghua National Laboratory for Information Science

and Technology, Beijing, China. (2)Department of Computer Science, Tsinghua

University, Beijing, China. (3)MOE Key Laboratory of Bioinformatics;

Bioinformatics Division and Center for Synthetic & Systems Biology, Tsinghua

National Laboratory for Information Science and Technology, Beijing, China.

ruijiang@tsinghua.edu.cn. (4)Department of Automation, Tsinghua University,

Beijing, China. ruijiang@tsinghua.edu.cn.

BACKGROUND: Whole exome sequencing (WES) has recently emerged as an effective

approach for identifying genetic variants underlying human diseases. However,

considerable time and labour is needed for careful investigation of candidate

variants. Although filtration based on population frequencies and functional

prediction scores could effectively remove common and neutral variants, hundreds

or even thousands of rare deleterious variants still remain. In addition, current

WES platforms also provide variant information in flanking noncoding regions,

such as promoters, introns and splice sites. Despite of being recognized to

harbour causal variants, these regions are usually ignored by current analysis

pipelines.

RESULTS: We present a novel computational method, called Glints, to overcome the

above limitations. Glints is capable of identifying disease-causing SNVs in both

coding and flanking noncoding regions from exome sequencing data. The principle

behind Glints is that disease-causing variants should manifest their effect at

both variant and gene levels. Specifically, Glints integrates 14 types of

functional scores, including predictions for both coding and noncoding variants,

and 9 types of association scores, which help identifying disease relevant genes.

We conducted a large-scale simulation studies based on 1000 Genomes Project data

and demonstrated the effectiveness of our method in both coding and flanking

noncoding regions. We also applied Glints in two real exome sequencing and

demonstrated its effectiveness for uncovering disease-causing SNVs. Both

standalone software and web server are available at our website

http://bioinfo.au.tsinghua.edu.cn/jianglab/glints .

CONCLUSIONS: Glints is effective for uncovering disease-causing SNVs in coding

and flanking noncoding regions, which is supported by both simulation and real

case studies. Glints is expected to be a useful tool for human genetics research

based on exome sequencing data.

DOI: 10.1186/s12859-016-1325-x

PMCID: PMC5260102

PMID: 28155632

35. BMC Syst Biol. 2016 Dec 23;10(Suppl 4):114. doi: 10.1186/s12918-016-0353-5.

Pretata: predicting TATA binding proteins with novel features and dimensionality

reduction strategy.

Zou Q(1), Wan S(1,)(2), Ju Y(3), Tang J(1,)(4), Zeng X(5).

Author information:

(1)School of Computer Science and Technology, Tianjin University, Tianjin, China.

(2)Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute

of Technology Shenzhen Graduate School, Shenzhen, Guangdong, 518055, China.

(3)School of Information Science and Engineering, Xiamen University, Xiamen,

China. (4)School of Computational Science and Engineering, University of South

Carolina, Columbia, USA. (5)School of Information Science and Engineering, Xiamen

University, Xiamen, China. xzeng@xmu.edu.cn.

BACKGROUND: It is necessary and essential to discovery protein function from the

novel primary sequences. Wet lab experimental procedures are not only

time-consuming, but also costly, so predicting protein structure and function

reliably based only on amino acid sequence has significant value. TATA-binding

protein (TBP) is a kind of DNA binding protein, which plays a key role in the

transcription regulation. Our study proposed an automatic approach for

identifying TATA-binding proteins efficiently, accurately, and conveniently. This

method would guide for the special protein identification with computational

intelligence strategies.

RESULTS: Firstly, we proposed novel fingerprint features for TBP based on pseudo

amino acid composition, physicochemical properties, and secondary structure.

Secondly, hierarchical features dimensionality reduction strategies were employed

to improve the performance furthermore. Currently, Pretata achieves 92.92%

TATA-binding protein prediction accuracy, which is better than all other existing

methods.

CONCLUSIONS: The experiments demonstrate that our method could greatly improve

the prediction accuracy and speed, thus allowing large-scale NGS data prediction

to be practical. A web server is developed to facilitate the other researchers,

which can be accessed at http://server.malab.cn/preTata/ .

DOI: 10.1186/s12918-016-0353-5

PMCID: PMC5259984

PMID: 28155714

36. Bioinformatics. 2016 Dec 22. pii: btw678. doi: 10.1093/bioinformatics/btw678.

[Epub ahead of print]

Improving protein disorder prediction by deep bidirectional long short-term

memory recurrent neural networks.

Hanson J(1), Yang Y(2), Paliwal K(1), Zhou Y(2).

Author information:

(1)Signal Processing Laboratory, Griffith University, Brisbane 4122, Australia.

(2)Institute for Glycomics, Griffith University, Gold Coast 4215, Australia.

MOTIVATION: Capturing long-range interactions between structural but not sequence

neighbors of proteins is a long-standing challenging problem in bioinformatics.

Recently, long short-term memory (LSTM) networks have significantly improved the

accuracy of speech and image classification problems by remembering useful past

information in long sequential events. Here, we have implemented deep

bidirectional LSTM recurrent neural networks in the problem of protein intrinsic

disorder prediction.

RESULTS: The new method, named SPOT-Disorder, has steadily improved over a

similar method using a traditional, window-based neural network (SPINE-D) in all

datasets tested without separate training on short and long disordered regions.

Independent tests on four other datasets including the datasets from critical

assessment of structure prediction (CASP) techniques and >10 000 annotated

proteins from MobiDB, confirmed SPOT-Disorder as one of the best methods in

disorder prediction. Moreover, initial studies indicate that the method is more

accurate in predicting functional sites in disordered regions. These results

highlight the usefulness combining LSTM with deep bidirectional recurrent neural

networks in capturing non-local, long-range interactions for bioinformatics

applications.

AVAILABILITY AND IMPLEMENTATION: SPOT-disorder is available as a web server and

as a standalone program at: http://sparks-lab.org/server/SPOT-disorder/index.php

CONTACT: j.hanson@griffith.edu.au or yuedong.yang@griffith.edu.au or

yaoqi.zhou@griffith.edu.auSupplementary information: Supplementary data is

available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw678

PMID: 28011771

37. Bioinformatics. 2016 Dec 22. pii: btw756. doi: 10.1093/bioinformatics/btw756.

[Epub ahead of print]

The PPI3D web server for searching, analyzing and modeling protein-protein

interactions in the context of 3D structures.

Dapkūnas J(1), Timinskas A(1), Olechnovič K(1,)(2), Margelevičius M(1), Dičiūnas

R(1), Venclovas Č(1).

Author information:

(1)Institute of Biotechnology, Vilnius University, Vilnius LT-10257, Lithuania.

(2)Faculty of Mathematics and Informatics, Vilnius University, Vilnius LT-03225,

Lithuania.

The PPI3D web server is focused on searching and analyzing the structural data on

protein-protein interactions. Reducing the data redundancy by clustering and

analyzing the properties of interaction interfaces using Voronoi tessellation

makes this software a highly effective tool for addressing different questions

related to protein interactions.AVAILABILITY AND IMPLEMENTATION: The server is

freely accessible at http://bioinformatics.lt/software/ppi3d/ CONTACT:

ceslovas.venclovas@bti.vu.ltSupplementary information: Supplementary data are

available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw756

PMID: 28011769

38. Bioinformatics. 2016 Dec 22. pii: btw783. doi: 10.1093/bioinformatics/btw783.

[Epub ahead of print]

RADER: a RApid DEcoy Retriever to facilitate decoy based assessment of virtual

screening.

Wang L(1), Pang X(1), Li Y(1), Zhang Z(2), Tan W(1).

Author information:

(1)Pre-Incubator for Innovative Drugs & Medicine, School of Bioscience and

Bioengineering, South China University of Technology, 382 East Circle at Higher

Education Mega Center, Guangzhou 510006, China. (2)Institute of Biomedical and

Pharmaceutical Sciences, Guangdong University of Technology, Guangzhou 510006,

China.

Evaluation of the capacity for separating actives from challenging decoys is a

crucial metric of performance related to molecular docking or a virtual screening

workflow. The Directory of Useful Decoys (DUD) and its enhanced version (DUD-E)

provide a benchmark for molecular docking, although they only contain a limited

set of decoys for limited targets. DecoyFinder was released to compensate the

limitations of DUD or DUD-E for building target-specific decoy sets. However,

desirable query template design, generation of multiple decoy sets of similar

quality, and computational speed remain bottlenecks, particularly when the

numbers of queried actives and retrieved decoys increases to hundreds or more.

Here, we developed a program suite called RApid DEcoy Retriever (RADER) to

facilitate the decoy-based assessment of virtual screening. This program adopts a

novel database-management regime that supports rapid and large-scale retrieval of

decoys, enables high portability of databases, and provides multifaceted options

for designing initial query templates from a large number of active ligands and

generating subtle decoy sets. RADER provides two operational modes: as a

command-line tool and on a web server. Validation of the performance and

efficiency of RADER was also conducted and is described.AVAILABILITY AND

IMPLEMENTATION: RADER web server and a local version are freely available at

http://rcidm.org/rader/ CONTACT: lingwang@scut.edu.cn or

went@scut.edu.cnSupplementary information: Supplementary data are available at

Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw783

PMID: 28011765

39. BMC Genomics. 2016 Dec 22;17(Suppl 13):1023. doi: 10.1186/s12864-016-3329-3.

LocExpress: a web server for efficiently estimating expression of novel

transcripts.

Hou M(1), Tian F(1,)(2), Jiang S(1), Kong L(1), Yang D(1), Gao G(3).

Author information:

(1)State Key Laboratory of Protein and Plant Gene Research, College of Life

Sciences, Center for Bioinformatics, Peking University, Beijing, 100871, People's

Republic of China. (2)Peking-Tsinghua Center for Life Sciences, Academy for

Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, People's

Republic of China. (3)State Key Laboratory of Protein and Plant Gene Research,

College of Life Sciences, Center for Bioinformatics, Peking University, Beijing,

100871, People's Republic of China. gaog@mail.cbi.pku.edu.cn.

BACKGROUND: The temporal and spatial-specific expression pattern of a transcript

in multiple tissues and cell types can indicate key clues about its function.

While several gene atlas available online as pre-computed databases for known

gene models, it's still challenging to get expression profile for previously

uncharacterized (i.e. novel) transcripts efficiently.

RESULTS: Here we developed LocExpress, a web server for efficiently estimating

expression of novel transcripts across multiple tissues and cell types in human

(20 normal tissues/cells types and 14 cell lines) as well as in mouse (24 normal

tissues/cell types and nine cell lines). As a wrapper to RNA-Seq quantification

algorithm, LocExpress efficiently reduces the time cost by making abundance

estimation calls increasingly within the minimum spanning bundle region of input

transcripts. For a given novel gene model, such local context-oriented strategy

allows LocExpress to estimate its FPKMs in hundreds of samples within minutes on

a standard Linux box, making an online web server possible.

CONCLUSIONS: To the best of our knowledge, LocExpress is the only web server to

provide nearly real-time expression estimation for novel transcripts in common

tissues and cell types. The server is publicly available at

http://loc-express.cbi.pku.edu.cn .

DOI: 10.1186/s12864-016-3329-3

PMCID: PMC5260097

PMID: 28155723

40. BMC Genomics. 2016 Dec 22;17(Suppl 13):1035. doi: 10.1186/s12864-016-3328-4.

ESAP plus: a web-based server for EST-SSR marker development.

Ponyared P(1), Ponsawat J(2), Tongsima S(3), Seresangtakul P(4), Akkasaeng C(5),

Tantisuwichwong N(6).

Author information:

(1)Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen,

40002, Thailand. (2)Department of Computer Engineering, Faculty of Engineering,

Khon Kaen University, Khon Kaen, 40002, Thailand. (3)National Center for Genetic

Engineering and Biotechnology (BIOTEC), Pathum Thani, 12120, Thailand.

(4)Department of Computer Science, Faculty of Science, Khon Kaen University, Khon

Kaen, 40002, Thailand. (5)Department of Plant Science and Agricultural Resources,

Faculty of Agriculture, Khon Kaen University, Khon Kaen, 40002, Thailand.

(6)Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen,

40002, Thailand. naptan@kku.ac.th.

BACKGROUND: Simple sequence repeats (SSRs) have become widely used as molecular

markers in plant genetic studies due to their abundance, high allelic variation

at each locus and simplicity to analyze using conventional PCR amplification. To

study plants with unknown genome sequence, SSR markers from Expressed Sequence

Tags (ESTs), which can be obtained from the plant mRNA (converted to cDNA), must

be utilized. With the advent of high-throughput sequencing technology, huge EST

sequence data have been generated and are now accessible from many public

databases. However, SSR marker identification from a large in-house or public EST

collection requires a computational pipeline that makes use of several standard

bioinformatic tools to design high quality EST-SSR primers. Some of these

computational tools are not users friendly and must be tightly integrated with

reference genomic databases.

RESULTS: A web-based bioinformatic pipeline, called EST Analysis Pipeline Plus

(ESAP Plus), was constructed for assisting researchers to develop SSR markers

from a large EST collection. ESAP Plus incorporates several bioinformatic scripts

and some useful standard software tools necessary for the four main procedures of

EST-SSR marker development, namely 1) pre-processing, 2) clustering and assembly,

3) SSR mining and 4) SSR primer design. The proposed pipeline also provides two

alternative steps for reducing EST redundancy and identifying SSR loci. Using

public sugarcane ESTs, ESAP Plus automatically executed the aforementioned

computational pipeline via a simple web user interface, which was implemented

using standard PHP, HTML, CSS and Java scripts. With ESAP Plus, users can upload

raw EST data and choose various filtering options and parameters to analyze each

of the four main procedures through this web interface. All input EST data and

their predicted SSR results will be stored in the ESAP Plus MySQL database. Users

will be notified via e-mail when the automatic process is completed and they can

download all the results through the web interface.

CONCLUSIONS: ESAP Plus is a comprehensive and convenient web-based bioinformatic

tool for SSR marker development. ESAP Plus offers all necessary EST-SSR

development processes with various adjustable options that users can easily use

to identify SSR markers from a large EST collection. With familiar web interface,

users can upload the raw EST using the data submission page and

visualize/download the corresponding EST-SSR information from within ESAP Plus.

ESAP Plus can handle considerably large EST datasets. This EST-SSR discovery tool

can be accessed directly from: http://gbp.kku.ac.th/esap\_plus/ .

DOI: 10.1186/s12864-016-3328-4

PMCID: PMC5260030

PMID: 28155670

41. BMC Genomics. 2016 Dec 22;17(Suppl 13):1027. doi: 10.1186/s12864-016-3326-6.

TEA: the epigenome platform for Arabidopsis methylome study.

Su SY(1,)(2,)(3), Chen SH(2), Lu IH(2), Chiang YS(4), Wang YB(2), Chen PY(5), Lin

CY(6,)(7,)(8).

Author information:

(1)Bioinformatics Program, Taiwan International Graduate Program, Institute of

Information Science, Academia Sinica, Taipei, Taiwan. (2)Institute of Information

Science, Academia Sinica, Taipei, Taiwan. (3)Institute of Biomedical Informatics,

National Yang-Ming University, Taipei, Taiwan. (4)Institute of Plant and

Microbial Biology, Academia Sinica, Taipei, Taiwan. (5)Institute of Plant and

Microbial Biology, Academia Sinica, Taipei, Taiwan. paoyang@gate.sinica.edu.tw.

(6)Institute of Information Science, Academia Sinica, Taipei, Taiwan.

cylin@iis.sinica.edu.tw. (7)Division of Biostatistics and Bioinformatics,

Institute of Population Health Sciences, National Health Research Institutes,

Zhunan, Miaoli, Taiwan. cylin@iis.sinica.edu.tw. (8)Institute of Fisheries

Science, College of Life Science, National Taiwan University, Taipei, Taiwan.

cylin@iis.sinica.edu.tw.

BACKGROUND: Bisulfite sequencing (BS-seq) has become a standard technology to

profile genome-wide DNA methylation at single-base resolution. It allows

researchers to conduct genome-wise cytosine methylation analyses on issues about

genomic imprinting, transcriptional regulation, cellular development and

differentiation. One single data from a BS-Seq experiment is resolved into many

features according to the sequence contexts, making methylome data analysis and

data visualization a complex task.

RESULTS: We developed a streamlined platform, TEA, for analyzing and visualizing

data from whole-genome BS-Seq (WGBS) experiments conducted in the model plant

Arabidopsis thaliana. To capture the essence of the genome methylation level and

to meet the efficiency for running online, we introduce a straightforward method

for measuring genome methylation in each sequence context by gene. The method is

scripted in Java to process BS-Seq mapping results. Through a simple data

uploading process, the TEA server deploys a web-based platform for deep analysis

by linking data to an updated Arabidopsis annotation database and toolkits.

CONCLUSIONS: TEA is an intuitive and efficient online platform for analyzing the

Arabidopsis genomic DNA methylation landscape. It provides several ways to help

users exploit WGBS data. TEA is freely accessible for academic users at:

http://tea.iis.sinica.edu.tw .

DOI: 10.1186/s12864-016-3326-6

PMCID: PMC5260138

PMID: 28155665

42. BMC Genomics. 2016 Dec 22;17(Suppl 13):1033. doi: 10.1186/s12864-016-3323-9.

An ensemble micro neural network approach for elucidating interactions between

zinc finger proteins and their target DNA.

Dutta S(1), Madan S(1), Parikh H(2), Sundar D(3).

Author information:

(1)Department of Biochemical Engineering and Biotechnology, DBT-AIST

International Laboratory for Advanced Biomedicine (DAILAB), Indian Institute of

Technology Delhi, New Delhi, 110016, India. (2)Department of Computer Science and

Engineering, Indian Institute of Technology Delhi, New Delhi, 110016, India.

(3)Department of Biochemical Engineering and Biotechnology, DBT-AIST

International Laboratory for Advanced Biomedicine (DAILAB), Indian Institute of

Technology Delhi, New Delhi, 110016, India. sundar@dbeb.iitd.ac.in.

BACKGROUND: The ability to engineer zinc finger proteins binding to a DNA

sequence of choice is essential for targeted genome editing to be possible.

Experimental techniques and molecular docking have been successful in predicting

protein-DNA interactions, however, they are highly time and resource intensive.

Here, we present a novel algorithm designed for high throughput prediction of

optimal zinc finger protein for 9 bp DNA sequences of choice. In accordance with

the principles of information theory, a subset identified by using K-means

clustering was used as a representative for the space of all possible 9 bp DNA

sequences. The modeling and simulation results assuming synergistic mode of

binding obtained from this subset were used to train an ensemble micro neural

network. Synergistic mode of binding is the closest to the DNA-protein binding

seen in nature, and gives much higher quality predictions, while the time and

resources increase exponentially in the trade off. Our algorithm is inspired from

an ensemble machine learning approach, and incorporates the predictions made by

100 parallel neural networks, each with a different hidden layer architecture

designed to pick up different features from the training dataset to predict

optimal zinc finger proteins for any 9 bp target DNA.

RESULTS: The model gave an accuracy of an average 83% sequence identity for the

testing dataset. The BLAST e-value are well within the statistical confidence

interval of E-05 for 100% of the testing samples. The geometric mean and median

value for the BLAST e-values were found to be 1.70E-12 and 7.00E-12 respectively.

For final validation of approach, we compared our predictions against optimal

ZFPs reported in literature for a set of experimentally studied DNA sequences.

The accuracy, as measured by the average string identity between our predictions

and the optimal zinc finger protein reported in literature for a 9 bp DNA target

was found to be as high as 81% for DNA targets with a consensus sequence

GCNGNNGCN reported in literature. Moreover, the average string identity of our

predictions for a catalogue of over 100 9 bp DNA for which the optimal zinc

finger protein has been reported in literature was found to be 71%.

CONCLUSIONS: Validation with experimental data shows that our tool is capable of

domain adaptation and thus scales well to datasets other than the training set

with high accuracy. As synergistic binding comes the closest to the ideal mode of

binding, our algorithm predicts biologically relevant results in sync with the

experimental data present in the literature. While there have been disjointed

attempts to approach this problem synergistically reported in literature, there

is no work covering the whole sample space. Our algorithm allows designing zinc

finger proteins for DNA targets of the user's choice, opening up new frontiers in

the field of targeted genome editing. This algorithm is also available as an easy

to use web server, ZifNN, at http://web.iitd.ac.in/~sundar/ZifNN/ .

DOI: 10.1186/s12864-016-3323-9

PMCID: PMC5260015

PMID: 28155662

43. BMC Genomics. 2016 Dec 22;17(Suppl 13):1037. doi: 10.1186/s12864-016-3324-8.

Exploiting the recognition code for elucidating the mechanism of zinc finger

protein-DNA interactions.

Dutta S(1), Madan S(1), Sundar D(2).

Author information:

(1)Department of Biochemical Engineering and Biotechnology, DBT-AIST

International Laboratory for Advanced Biomedicine (DAILAB), Indian Institute of

Technology Delhi, New Delhi, 110016, India. (2)Department of Biochemical

Engineering and Biotechnology, DBT-AIST International Laboratory for Advanced

Biomedicine (DAILAB), Indian Institute of Technology Delhi, New Delhi, 110016,

India. sundar@dbeb.iitd.ac.in.

BACKGROUND: Engineering zinc finger protein motifs for specific binding to

double-stranded DNA is critical for targeted genome editing. Most existing tools

for predicting DNA-binding specificity in zinc fingers are trained on data

obtained from naturally occurring proteins, thereby skewing the predictions.

Moreover, these mostly neglect the cooperativity exhibited by zinc fingers.

METHODS: Here, we present an ab-initio method that is based on mutation of the

key α-helical residues of individual fingers of the parent template for Zif-268

and its consensus sequence (PDB ID: 1AAY). In an attempt to elucidate the

mechanism of zinc finger protein-DNA interactions, we evaluated and compared

three approaches, differing in the amino acid mutations introduced in the Zif-268

parent template, and the mode of binding they try to mimic, i.e., modular and

synergistic mode of binding.

RESULTS: Comparative evaluation of the three strategies reveals that the

synergistic mode of binding appears to mimic the ideal mechanism of DNA-zinc

finger protein binding. Analysis of the predictions made by all three strategies

indicate strong dependence of zinc finger binding specificity on the amino acid

propensity and the position of a 3-bp DNA sub-site in the target DNA sequence.

Moreover, the binding affinity of the individual zinc fingers was found to

increase in the order Finger 1 < Finger 2 < Finger 3, thus confirming the

cooperative effect.

CONCLUSIONS: Our analysis offers novel insights into the prediction of ZFPs for

target DNA sequences and the approaches have been made available as an easy to

use web server at http://web.iitd.ac.in/~sundar/zifpredict\_ihbe.

DOI: 10.1186/s12864-016-3324-8

PMCID: PMC5260074

PMID: 28155654

44. J Cheminform. 2016 Dec 20;8:72. doi: 10.1186/s13321-016-0185-8. eCollection 2016.

osFP: a web server for predicting the oligomeric states of fluorescent proteins.

Simeon S(1), Shoombuatong W(1), Anuwongcharoen N(1), Preeyanon L(2),

Prachayasittikul V(2), Wikberg JE(3), Nantasenamat C(1).

Author information:

(1)Center of Data Mining and Biomedical Informatics, Faculty of Medical

Technology, Mahidol University, Bangkok, 10700 Thailand. (2)Department of

Community Medical Technology, Faculty of Medical Technology, Mahidol University,

Bangkok, 10700 Thailand. (3)Department of Pharmaceutical Biosciences, Uppsala

University, 751 24 Uppsala, Sweden.

BACKGROUND: Currently, monomeric fluorescent proteins (FP) are ideal markers for

protein tagging. The prediction of oligomeric states is helpful for enhancing

live biomedical imaging. Computational prediction of FP oligomeric states can

accelerate the effort of protein engineering efforts of creating monomeric FPs.

To the best of our knowledge, this study represents the first computational model

for predicting and analyzing FP oligomerization directly from the amino acid

sequence.

RESULTS: After data curation, an exhaustive data set consisting of 397

non-redundant FP oligomeric states was compiled from the literature. Results from

benchmarking of the protein descriptors revealed that the model built with amino

acid composition descriptors was the top performing model with accuracy,

sensitivity and specificity in excess of 80% and MCC greater than 0.6 for all

three data subsets (e.g. training, tenfold cross-validation and external sets).

The model provided insights on the important residues governing the

oligomerization of FP. To maximize the benefit of the generated predictive model,

it was implemented as a web server under the R programming environment.

CONCLUSION: osFP affords a user-friendly interface that can be used to predict

the oligomeric state of FP using the protein sequence. The advantage of osFP is

that it is platform-independent meaning that it can be accessed via a web browser

on any operating system and device. osFP is freely accessible at

http://codes.bio/osfp/ while the source code and data set is provided on GitHub

at https://github.com/chaninn/osFP/.Graphical Abstract.

DOI: 10.1186/s13321-016-0185-8

PMCID: PMC5167684

PMID: 28053671

45. Retrovirology. 2016 Dec 20;13(1):85. doi: 10.1186/s12977-016-0320-7.

A genotypic method for determining HIV-2 coreceptor usage enables epidemiological

studies and clinical decision support.

Döring M(1), Borrego P(2,)(3), Büch J(4), Martins A(2), Friedrich G(4), Camacho

RJ(5), Eberle J(6), Kaiser R(7), Lengauer T(4), Taveira N(2,)(8), Pfeifer N(9).

Author information:

(1)Department for Computational Biology and Applied Algorithmics, Max Planck

Institute for Informatics, Saarland Informatics Campus, Campus E 1 4, 66123,

Saarbrücken, Germany. mdoering@mpi-inf.mpg.de. (2)Research Institute for

Medicines (iMed.ULisboa), Faculty of Pharmacy, University of Lisbon, Av.

Professor Gama Pinto, 1649-003, Lisbon, Portugal. (3)Centro de Administração e

Políticas Públicas (CAPP), Instituto Superior de Ciências Sociais e Políticas

(ISCSP), University of Lisbon, Rua Almerindo Lessa, 1300-663, Lisbon, Portugal.

(4)Department for Computational Biology and Applied Algorithmics, Max Planck

Institute for Informatics, Saarland Informatics Campus, Campus E 1 4, 66123,

Saarbrücken, Germany. (5)Rega Institute for Medical Research, Clinical and

Epidemiological Virology, Department of Microbiology and Immunology, KU

Leuven-University of Leuven, Minderbroedersstraat 10, 3000, Louvain, Belgium.

(6)Department of Virology, Max von Pettenkofer-Institut,

Ludwig-Maximilians-University, Pettenkoferstraße 9a, 80336, Munich, Germany.

(7)Institute for Virology, University of Cologne, Fürst-Pückler-Str. 56, 50935,

Cologne, Germany. (8)Instituto Superior de Ciências da Saúde Egas Moniz (ISCSEM),

Campus Universitário, Quinta da Granja, Monte de Caparica, 2829-511, Caparica,

Portugal. (9)Department for Computational Biology and Applied Algorithmics, Max

Planck Institute for Informatics, Saarland Informatics Campus, Campus E 1 4,

66123, Saarbrücken, Germany. npfeifer@mpi-inf.mpg.de.

BACKGROUND: CCR5-coreceptor antagonists can be used for treating HIV-2 infected

individuals. Before initiating treatment with coreceptor antagonists, viral

coreceptor usage should be determined to ensure that the virus can use only the

CCR5 coreceptor (R5) and cannot evade the drug by using the CXCR4 coreceptor

(X4-capable). However, until now, no online tool for the genotypic identification

of HIV-2 coreceptor usage had been available. Furthermore, there is a lack of

knowledge on the determinants of HIV-2 coreceptor usage. Therefore, we developed

a data-driven web service for the prediction of HIV-2 coreceptor usage from the

V3 loop of the HIV-2 glycoprotein and used the tool to identify novel

discriminatory features of X4-capable variants.

RESULTS: Using 10 runs of tenfold cross validation, we selected a linear support

vector machine (SVM) as the model for geno2pheno[coreceptor-hiv2], because it

outperformed the other SVMs with an area under the ROC curve (AUC) of 0.95. We

found that SVMs were highly accurate in identifying HIV-2 coreceptor usage,

attaining sensitivities of 73.5% and specificities of 96% during tenfold nested

cross validation. The predictive performance of SVMs was not significantly

different (p value 0.37) from an existing rules-based approach. Moreover,

geno2pheno[coreceptor-hiv2] achieved a predictive accuracy of 100% and

outperformed the existing approach on an independent data set containing nine new

isolates with corresponding phenotypic measurements of coreceptor usage.

geno2pheno[coreceptor-hiv2] could not only reproduce the established markers of

CXCR4-usage, but also revealed novel markers: the substitutions 27K, 15G, and 8S

were significantly predictive of CXCR4 usage. Furthermore, SVMs trained on the

amino-acid sequences of the V1 and V2 loops were also quite accurate in

predicting coreceptor usage (AUCs of 0.84 and 0.65, respectively).

CONCLUSIONS: In this study, we developed geno2pheno[coreceptor-hiv2], the first

online tool for the prediction of HIV-2 coreceptor usage from the V3 loop. Using

our method, we identified novel amino-acid markers of X4-capable variants in the

V3 loop and found that HIV-2 coreceptor usage is also influenced by the V1/V2

region. The tool can aid clinicians in deciding whether coreceptor antagonists

such as maraviroc are a treatment option and enables epidemiological studies

investigating HIV-2 coreceptor usage. geno2pheno[coreceptor-hiv2] is freely

available at http://coreceptor-hiv2.geno2pheno.org .

DOI: 10.1186/s12977-016-0320-7

PMCID: PMC5168878

PMID: 27998283

46. Bioinformatics. 2016 Dec 19. pii: btw702. [Epub ahead of print]

GENIUS: web server to predict local gene networks and key genes for biological

functions.

Puelma T(1,)(2), Araus V(1), Canales J(1), Vidal EA(1), Cabello JM(1), Soto A(2),

Gutiérrez RA(1).

Author information:

(1)Millennium Nucleus Center for Plant Systems and Synthetic Biology, FONDAP

Center for Genome Regulation, Departamento de Genética Molecular y Microbiología,

Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile,

Santiago 8331150, Chile. (2)Millennium Nucleus Center for Plant Systems and

Synthetic Biology, Departamento de Ciencias de la Computación, Facultad de

Ingeniería, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile.

GENIUS is a user-friendly web server that uses a novel machine learning algorithm

to infer functional gene networks focused on specific genes and experimental

conditions that are relevant to biological functions of interest. These functions

may have different levels of complexity, from specific biological processes to

complex traits that involve several interacting processes. GENIUS also enriches

the network with new genes related to the biological function of interest, with

accuracies comparable to highly discriminative Support Vector Machine

methods.AVAILABILITY AND IMPLEMENTATION: GENIUS currently supports eight model

organisms and is freely available for public use at

http://networks.bio.puc.cl/genius CONTACT: genius.psbl@gmail.comSupplementary

information: Supplementary data are available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw702

PMID: 27993775

47. Bioinformatics. 2016 Dec 15;32(24):3745-3752. Epub 2016 Aug 26.

Imbalanced multi-label learning for identifying antimicrobial peptides and their

functional types.

Lin W(1,)(2), Xu D(2).

Author information:

(1)nformation Engineering School, Jingdezhen Ceramic Institute, Jingdezhen

333406, China. (2)Department of Computer Science and Christopher S. Bond Life

Sciences Center, University of Missouri, Columbia, MO 65211, USA.

MOTIVATION: With the rapid increase of infection resistance to antibiotics, it is

urgent to find novel infection therapeutics. In recent years, antimicrobial

peptides (AMPs) have been utilized as potential alternatives for infection

therapeutics. AMPs are key components of the innate immune system and can protect

the host from various pathogenic bacteria. Identifying AMPs and their functional

types has led to many studies, and various predictors using machine learning have

been developed. However, there is room for improvement; in particular, no

predictor takes into account the lack of balance among different functional AMPs.

RESULTS: In this paper, a new synthetic minority over-sampling technique on

imbalanced and multi-label datasets, referred to as ML-SMOTE, was designed for

processing and identifying AMPs' functional families. A novel multi-label

classifier, MLAMP, was also developed using ML-SMOTE and grey pseudo amino acid

composition. The classifier obtained 0.4846 subset accuracy and 0.16 hamming

loss.

AVAILABILITY AND IMPLEMENTATION: A user-friendly web-server for MLAMP was

established at http://www.jci-bioinfo.cn/MLAMP CONTACTS: linweizhong@jci.edu.cn

or xudong@missouri.edu.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw560

PMCID: PMC5167070 [Available on 2017-12-15]

PMID: 27565585

48. Front Microbiol. 2016 Dec 15;7:2010. doi: 10.3389/fmicb.2016.02010. eCollection

2016.

Construction of a Pan-Genome Allele Database of Salmonella enterica Serovar

Enteritidis for Molecular Subtyping and Disease Cluster Identification.

Liu YY(1), Chen CC(2), Chiou CS(1).

Author information:

(1)Central Regional Laboratory, Center for Diagnostics and Vaccine Development,

Centers for Disease Control Taichung, Taiwan. (2)Institute of Medical Science and

Technology, National Sun Yat-sen UniversityKaohsiung, Taiwan; Medical Science and

Technology Center, National Sun Yat-sen UniversityKaohsiung, Taiwan.

We built a pan-genome allele database with 395 genomes of Salmonella enterica

serovar Enteritidis and developed computer tools for analysis of whole genome

sequencing (WGS) data of bacterial isolates for disease cluster identification. A

web server (http://wgmlst.imst.nsysu.edu.tw) was set up with the database and the

tools, allowing users to upload WGS data to generate whole genome multilocus

sequence typing (wgMLST) profiles and to perform cluster analysis of wgMLST

profiles. The usefulness of the database in disease cluster identification was

demonstrated by analyzing a panel of genomes from 55 epidemiologically

well-defined S. Enteritidis isolates provided by the Minnesota Department of

Health. The wgMLST-based cluster analysis revealed distinct clades that were

concordant with the epidemiologically defined outbreaks. Thus, using a common

pan-genome allele database, wgMLST can be a promising WGS-based subtyping

approach for disease surveillance and outbreak investigation across laboratories.

DOI: 10.3389/fmicb.2016.02010

PMCID: PMC5156723

PMID: 28018331

49. Oncotarget. 2016 Dec 13;7(50):82440-82449. doi: 10.18632/oncotarget.12517.

Identification of self-interacting proteins by exploring evolutionary information

embedded in PSI-BLAST-constructed position specific scoring matrix.

An JY(1), You ZH(2), Chen X(3), Huang DS(4), Li ZW(1), Liu G(5), Wang Y(1).

Author information:

(1)School of Computer Science and Technology, China University of Mining and

Technology, Xuzhou 21116, China. (2)Xinjiang Technical Institute of Physics and

Chemistry, Chinese Academy of Science, Urumqi 830011, China. (3)School of

Information and Electrical Engineering, China University of Mining and

Technology, Xuzhou 221116, China. (4)School of Electronics and Information

Engineering, Tongji University, Shanghai 201804, China. (5)College of Computer

Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong

518060, China.

Self-interacting Proteins (SIPs) play an essential role in a wide range of

biological processes, such as gene expression regulation, signal transduction,

enzyme activation and immune response. Because of the limitations for

experimental self-interaction proteins identification, developing an effective

computational method based on protein sequence to detect SIPs is much important.

In the study, we proposed a novel computational approach called RVMBIGP that

combines the Relevance Vector Machine (RVM) model and Bi-gram probability (BIGP)

to predict SIPs based on protein sequence. The proposed prediction model includes

as following steps: (1) an effective feature extraction method named BIGP is used

to represent protein sequences on Position Specific Scoring Matrix (PSSM); (2)

Principal Component Analysis (PCA) method is employed for integrating the useful

information and reducing the influence of noise; (3) the robust classifier

Relevance Vector Machine (RVM) is used to carry out classification. When

performed on yeast and human datasets, the proposed RVMBIGP model can achieve

very high accuracies of 95.48% and 98.80%, respectively. The experimental results

show that our proposed method is very promising and may provide a cost-effective

alternative for SIPs identification. In addition, to facilitate extensive studies

for future proteomics research, the RVMBIGP server is freely available for

academic use at http://219.219.62.123:8888/RVMBIGP.

DOI: 10.18632/oncotarget.12517

PMID: 27732957

50. Sci Rep. 2016 Dec 13;6:38881. doi: 10.1038/srep38881.

Performance Evaluation and Online Realization of Data-driven Normalization

Methods Used in LC/MS based Untargeted Metabolomics Analysis.

Li B(1), Tang J(1), Yang Q(1), Cui X(1), Li S(1), Chen S(2), Cao Q(1), Xue W(1),

Chen N(1), Zhu F(1).

Author information:

(1)Innovative Drug Research and Bioinformatics Group, Innovative Drug Research

Centre and School of Pharmaceutical Sciences, Chongqing University, Chongqing

401331, China. (2)College of Mathematics and Statistics, Chongqing University,

Chongqing 401331, China.

In untargeted metabolomics analysis, several factors (e.g., unwanted experimental

&biological variations and technical errors) may hamper the identification of

differential metabolic features, which requires the data-driven normalization

approaches before feature selection. So far, ≥16 normalization methods have been

widely applied for processing the LC/MS based metabolomics data. However, the

performance and the sample size dependence of those methods have not yet been

exhaustively compared and no online tool for comparatively and comprehensively

evaluating the performance of all 16 normalization methods has been provided. In

this study, a comprehensive comparison on these methods was conducted. As a

result, 16 methods were categorized into three groups based on their

normalization performances across various sample sizes. The VSN, the Log

Transformation and the PQN were identified as methods of the best normalization

performance, while the Contrast consistently underperformed across all

sub-datasets of different benchmark data. Moreover, an interactive web tool

comprehensively evaluating the performance of 16 methods specifically for

normalizing LC/MS based metabolomics data was constructed and hosted at

http://server.idrb.cqu.edu.cn/MetaPre/. In summary, this study could serve as a

useful guidance to the selection of suitable normalization methods in analyzing

the LC/MS based metabolomics data.

DOI: 10.1038/srep38881

PMCID: PMC5153651

PMID: 27958387

51. F1000Res. 2016 Dec 12;5. pii: ELIXIR-2841. doi: 10.12688/f1000research.10221.1.

eCollection 2016.

Integration of EGA secure data access into Galaxy.

Hoogstrate Y(1), Zhang C(2), Senf A(3), Bijlard J(4), Hiltemann S(1), van

Enckevort D(5), Repo S(6), Heringa J(2), Jenster G(7), J A Fijneman R(8), Boiten

JW(9), A Meijer G(8), Stubbs A(1), Rambla J(10), Spalding D(3), Abeln S(2).

Author information:

(1)Department of Bioinformatics, ErasmusMC Rotterdam, Rotterdam, Netherlands.

(2)Department of Computer Science, Vrije Universiteit, Amsterdam, Netherlands.

(3)European Molecular Biology Laboratory, European Bioinformatics Institute,

Wellcome Genome Campus, Hinxton, UK. (4)The Hyve, Utrecht, Netherlands.

(5)Genomics Coordination Centre, UMC Groningen, Groningen, Netherlands. (6)ELIXIR

Hub, Wellcome Genome Campus, Hinxton, UK. (7)Department of Urology, ErasmusMC

Rotterdam, Rotterdam, Netherlands. (8)Diagnostic Oncology, Netherlands Cancer

Institute, Amsterdam, Netherlands. (9)Lygature, Utrecht, Netherlands. (10)Centre

for Genomic Regulation, Parc de Recerca Biomédica de Barcelona, Barcelona, Spain.

High-throughput molecular profiling techniques are routinely generating vast

amounts of data for translational medicine studies. Secure access controlled

systems are needed to manage, store, transfer and distribute these data due to

its personally identifiable nature. The European Genome-phenome Archive (EGA) was

created to facilitate access and management to long-term archival of

bio-molecular data. Each data provider is responsible for ensuring a Data Access

Committee is in place to grant access to data stored in the EGA. Moreover, the

transfer of data during upload and download is encrypted. ELIXIR, a European

research infrastructure for life-science data, initiated a project (2016 Human

Data Implementation Study) to understand and document the ELIXIR requirements for

secure management of controlled-access data. As part of this project, a full

ecosystem was designed to connect archived raw experimental molecular profiling

data with interpreted data and the computational workflows, using the CTMM

Translational Research IT (CTMM-TraIT) infrastructure http://www.ctmm-trait.nl as

an example. Here we present the first outcomes of this project, a framework to

enable the download of EGA data to a Galaxy server in a secure way. Galaxy

provides an intuitive user interface for molecular biologists and

bioinformaticians to run and design data analysis workflows. More specifically,

we developed a tool -- ega\_download\_streamer - that can download data securely

from EGA into a Galaxy server, which can subsequently be further processed. This

tool will allow a user within the browser to run an entire analysis containing

sensitive data from EGA, and to make this analysis available for other

researchers in a reproducible manner, as shown with a proof of concept study.

 The tool ega\_download\_streamer is available in the Galaxy tool shed:

https://toolshed.g2.bx.psu.edu/view/yhoogstrate/ega\_download\_streamer.

DOI: 10.12688/f1000research.10221.1

PMCID: PMC5302147

PMID: 28232859

Conflict of interest statement: Competing interests: No competing interests were

disclosed.

52. G3 (Bethesda). 2016 Dec 7;6(12):3927-3939. doi: 10.1534/g3.116.034744.

Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using

Genome-Scale Data.

Shen XX(1), Zhou X(1), Kominek J(2), Kurtzman CP(3), Hittinger CT(4), Rokas A(5).

Author information:

(1)Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee

37235. (2)Laboratory of Genetics, Genome Center of Wisconsin, DOE Great Lakes

Bioenergy Research Center, Wisconsin Energy Institute, J. F. Crow Institute for

the Study of Evolution, University of Wisconsin-Madison, Madison, Wisconsin

53706. (3)Mycotoxin Prevention and Applied Microbiology Research Unit, National

Center for Agricultural Utilization Research, Agricultural Research Service, U.S.

Department of Agriculture, Peoria, Illinois 61604 antonis.rokas@vanderbilt.edu.

(4)Laboratory of Genetics, Genome Center of Wisconsin, DOE Great Lakes Bioenergy

Research Center, Wisconsin Energy Institute, J. F. Crow Institute for the Study

of Evolution, University of Wisconsin-Madison, Madison, Wisconsin 53706

antonis.rokas@vanderbilt.edu. (5)Department of Biological Sciences, Vanderbilt

University, Nashville, Tennessee 37235 antonis.rokas@vanderbilt.edu.

Understanding the phylogenetic relationships among the yeasts of the subphylum

Saccharomycotina is a prerequisite for understanding the evolution of their

metabolisms and ecological lifestyles. In the last two decades, the use of rDNA

and multilocus data sets has greatly advanced our understanding of the yeast

phylogeny, but many deep relationships remain unsupported. In contrast,

phylogenomic analyses have involved relatively few taxa and lineages that were

often selected with limited considerations for covering the breadth of yeast

biodiversity. Here we used genome sequence data from 86 publicly available yeast

genomes representing nine of the 11 known major lineages and 10 nonyeast fungal

outgroups to generate a 1233-gene, 96-taxon data matrix. Species phylogenies

reconstructed using two different methods (concatenation and coalescence) and two

data matrices (amino acids or the first two codon positions) yielded identical

and highly supported relationships between the nine major lineages. Aside from

the lineage comprised by the family Pichiaceae, all other lineages were

monophyletic. Most interrelationships among yeast species were robust across the

two methods and data matrices. However, eight of the 93 internodes conflicted

between analyses or data sets, including the placements of: the clade defined by

species that have reassigned the CUG codon to encode serine, instead of leucine;

the clade defined by a whole genome duplication; and the species Ascoidea

rubescens These phylogenomic analyses provide a robust roadmap for future

comparative work across the yeast subphylum in the disciplines of taxonomy,

molecular genetics, evolutionary biology, ecology, and biotechnology. To further

this end, we have also provided a BLAST server to query the 86 Saccharomycotina

genomes, which can be found at http://y1000plus.org/blast.

Copyright © 2016 Shen et al.

DOI: 10.1534/g3.116.034744

PMCID: PMC5144963

PMID: 27672114

53. Sci Rep. 2016 Dec 6;6:38367. doi: 10.1038/srep38367.

Resistance gene identification from Larimichthys crocea with machine learning

techniques.

Cai Y(1,)(2), Liao Z(3), Ju Y(1), Liu J(4), Mao Y(2,)(5), Liu X(1,)(2).

Author information:

(1)School of Information Science and Technology, Xiamen University, Xiamen,

Fujian 361005, China. (2)State Key Laboratory of Large Yellow Croaker Breeding,

Ningde Fufa Fisheries Company Limited, Ningde, 352000, China. (3)Department of

Biochemistry and Molecular Biology, School of Basic Medical Sciences, Fujian

Medical University, Fuzhou, 350122, China. (4)School of Aerospace Engineering,

Xiamen University, Xiamen, Fujian 361005, China. (5)College of Ocean and Earth

Sciences, Xiamen University, Xiamen, 361102, China.

The research on resistance genes (R-gene) plays a vital role in bioinformatics as

it has the capability of coping with adverse changes in the external environment,

which can form the corresponding resistance protein by transcription and

translation. It is meaningful to identify and predict R-gene of Larimichthys

crocea (L.Crocea). It is friendly for breeding and the marine environment as

well. Large amounts of L.Crocea's immune mechanisms have been explored by

biological methods. However, much about them is still unclear. In order to break

the limited understanding of the L.Crocea's immune mechanisms and to detect new

R-gene and R-gene-like genes, this paper came up with a more useful combination

prediction method, which is to extract and classify the feature of available

genomic data by machine learning. The effectiveness of feature extraction and

classification methods to identify potential novel R-gene was evaluated, and

different statistical analyzes were utilized to explore the reliability of

prediction method, which can help us further understand the immune mechanisms of

L.Crocea against pathogens. In this paper, a webserver called LCRG-Pred is

available at http://server.malab.cn/rg\_lc/.

DOI: 10.1038/srep38367

PMCID: PMC5138596

PMID: 27922074

54. J Proteome Res. 2016 Dec 2;15(12):4755-4762. Epub 2016 Nov 3.

Ensemble Linear Neighborhood Propagation for Predicting Subchloroplast

Localization of Multi-Location Proteins.

Wan S(1), Mak MW(1), Kung SY(2).

Author information:

(1)Department of Electronic and Information Engineering, The Hong Kong

Polytechnic University , Hong Kong SAR, China. (2)Department of Electrical

Engineering, Princeton University , New Jersey 08540, United States.

In the postgenomic era, the number of unreviewed protein sequences is remarkably

larger and grows tremendously faster than that of reviewed ones. However,

existing methods for protein subchloroplast localization often ignore the

information from these unlabeled proteins. This paper proposes a multi-label

predictor based on ensemble linear neighborhood propagation (LNP), namely,

LNP-Chlo, which leverages hybrid sequence-based feature information from both

labeled and unlabeled proteins for predicting localization of both single- and

multi-label chloroplast proteins. Experimental results on a stringent benchmark

dataset and a novel independent dataset suggest that LNP-Chlo performs at least

6% (absolute) better than state-of-the-art predictors. This paper also

demonstrates that ensemble LNP significantly outperforms LNP based on individual

features. For readers' convenience, the online Web server LNP-Chlo is freely

available at http://bioinfo.eie.polyu.edu.hk/LNPChloServer/ .

DOI: 10.1021/acs.jproteome.6b00686

PMID: 27766879

55. Sci Rep. 2016 Dec 2;6:38318. doi: 10.1038/srep38318.

Mal-Lys: prediction of lysine malonylation sites in proteins integrated

sequence-based features with mRMR feature selection.

Xu Y(1), Ding YX(1), Ding J(1), Wu LY(2), Xue Y(3).

Author information:

(1)Department of Information and Computer Science, University of Science and

Technology Beijing, Beijing 100083, China. (2)Institute of Applied Mathematics,

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing

100190, China. (3)Department of Biomedical Engineering, College of Life Science

and Technology, Huazhong University of Science and Technology, Wuhan, Hubei

430074, China.

Lysine malonylation is an important post-translational modification (PTM) in

proteins, and has been characterized to be associated with diseases. However,

identifying malonyllysine sites still remains to be a great challenge due to the

labor-intensive and time-consuming experiments. In view of this situation, the

establishment of a useful computational method and the development of an

efficient predictor are highly desired. In this study, a predictor Mal-Lys which

incorporated residue sequence order information, position-specific amino acid

propensity and physicochemical properties was proposed. A feature selection

method of minimum Redundancy Maximum Relevance (mRMR) was used to select optimal

ones from the whole features. With the leave-one-out validation, the value of the

area under the curve (AUC) was calculated as 0.8143, whereas 6-, 8- and 10-fold

cross-validations had similar AUC values which showed the robustness of the

predictor Mal-Lys. The predictor also showed satisfying performance in the

experimental data from the UniProt database. Meanwhile, a user-friendly

web-server for Mal-Lys is accessible at http://app.aporc.org/Mal-Lys/.

DOI: 10.1038/srep38318

PMCID: PMC5133563

PMID: 27910954

56. Bioinformatics. 2016 Dec 1;32(23):3584-3592. Epub 2016 Aug 11.

ChemTreeMap: an interactive map of biochemical similarity in molecular datasets.

Lu J(1), Carlson HA(1,)(2).

Author information:

(1)Department of Computational Medicine and Bioinformatics. (2)Department of

Medicinal Chemistry, University of Michigan, Ann Arbor, MI, USA.

MOTIVATION: What if you could explain complex chemistry in a simple tree and

share that data online with your collaborators? Computational biology often

incorporates diverse chemical data to probe a biological question, but the

existing tools for chemical data are ill-suited for the very large datasets

inherent to bioinformatics. Furthermore, existing visualization methods often

require an expert chemist to interpret the patterns. Biologists need an

interactive tool for visualizing chemical information in an intuitive, accessible

way that facilitates its integration into today's team-based biological research.

RESULTS: ChemTreeMap is an interactive, bioinformatics tool designed to explore

chemical space and mine the relationships between chemical structure, molecular

properties, and biological activity. ChemTreeMap synergistically combines

extended connectivity fingerprints and a neighbor-joining algorithm to produce a

hierarchical tree with branch lengths proportional to molecular similarity.

Compound properties are shown by leaf color, size and outline to yield a

user-defined visualization of the tree. Two representative analyses are included

to demonstrate ChemTreeMap's capabilities and utility: assessing dataset overlap

and mining structure-activity relationships.

AVAILABILITY AND IMPLEMENTATION: The examples from this paper may be accessed at

http://ajing.github.io/ChemTreeMap/ Code for the server and client are available

in the Supplementary Information, at the aforementioned github site, and on

Docker Hub (https://hub.docker.com) with the nametag ajing/chemtreemap.

CONTACT: carlsonh@umich.eduSupplementary information: Supplementary data are

available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw523

PMCID: PMC5181537 [Available on 2017-12-01]

PMID: 27515740

57. Bioinformatics. 2016 Dec 1;32(23):3676-3678. Epub 2016 Aug 8.

PRODIGY: a web server for predicting the binding affinity of protein-protein

complexes.

Xue LC(1), Rodrigues JP(1), Kastritis PL(1), Bonvin AM(1), Vangone A(1).

Author information:

(1)Computational Structural Biology Group, Bijvoet Center for Biomolecular

Research, Faculty of Science - Department of Chemistry, Utrecht University,

3584CH Utrecht, The Netherlands.

Gaining insights into the structural determinants of protein-protein interactions

holds the key for a deeper understanding of biological functions, diseases and

development of therapeutics. An important aspect of this is the ability to

accurately predict the binding strength for a given protein-protein complex. Here

we present PROtein binDIng enerGY prediction (PRODIGY), a web server to predict

the binding affinity of protein-protein complexes from their 3D structure. The

PRODIGY server implements our simple but highly effective predictive model based

on intermolecular contacts and properties derived from non-interface

surface.AVAILABILITY AND IMPLEMENTATION: PRODIGY is freely available at:

http://milou.science.uu.nl/services/PRODIGY CONTACT: a.m.j.j.bonvin@uu.nl,

a.vangone@uu.nl.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw514

PMID: 27503228

58. Bioinformatics. 2016 Dec 1;32(23):3673-3675. Epub 2016 Aug 6.

WebSTAR3D: a web server for RNA 3D structural alignment.

Holzhauser E(1), Ge P(1), Zhang S(1).

Author information:

(1)Department of Computer Science, University of Central Florida, Orlando, FL

32816, USA.

The WebSTAR3D web server is a user-friendly online interface for the alignment of

RNA 3D structures. The website takes as input two files, each of which can be in

either PDB or mmCIF format, containing the desired structures to align, via a PDB

code or user upload. In return, the user is presented with a visualization of the

aligned structures in Jmol or JSmol, along with the corresponding sequence

alignment, and the option to download the nucleotide mapping of the structures

and a PDB file containing the aligned, superimposed structures.AVAILABILITY AND

IMPLEMENTATION: The WebSTAR3D is available at http://rna.ucf.edu/WebSTAR3D

CONTACT: shzhang@cs.ucf.edu.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw502

PMCID: PMC5181533 [Available on 2017-12-01]

PMID: 27497443

59. Comput Biol Med. 2016 Dec 1;79:30-35. doi: 10.1016/j.compbiomed.2016.10.003. Epub

2016 Oct 4.

Computational identification of non-synonymous polymorphisms within regions

corresponding to protein interaction sites.

Skrlj B(1), Kunej T(2).

Author information:

(1)Department of Animal Science, Biotechnical Faculty, University of Ljubljana,

Groblje 3, Slovenia. Electronic address: skrljblaz@gmail.com. (2)Department of

Animal Science, Biotechnical Faculty, University of Ljubljana, Groblje 3,

Slovenia. Electronic address: tanja.kunej@bf.uni-lj.si.

BACKGROUND: Protein-protein interactions (PPI) play an important role in function

of all organisms and enable understanding of underlying metabolic processes.

Computational predictions of PPIs are an important aspect in proteomics, as

experimental methods may result in high degree of false positive results and are

more expensive. Although there are many databases collecting predicted PPIs,

exploration of genetics information underlying PPI interactions has not been

investigated thoroughly. The aim of the present study was to identify genomic

locations corresponding to regions involved in predicted PPIs and to collect

non-synonymous polymorphisms (nsSNPs) located within those regions; which we

termed PPI-SNPs.

METHODS: Predicted PPIs were obtained from PiSITE database

(http://pisite.hgc.jp). Non-synonymous SNPs mapped on protein structural data

(PDBs) were obtained from the UCSC server. Polymorphism locations on protein

structures were mapped to predicted PPI regions. DAVID tool was used for pathway

enrichment and gene cluster analysis (https://david.ncifcrf.gov/).

RESULTS: We collected 544 polymorphisms located within predicted PPI sites that

map to 197 genes. We identified 9 SNPs, previously associated with diseases, but

not yet associated with PPI sites. We also found examples in which polymorphisms

located within predicted PPI regions are also occurring within previously

experimentally validated PPIs and within experimentally determined functional

domains.

CONCLUSIONS: Our study provides the first catalog of nsSNPs located within

predicted PPIs. These prioritized SNPs present the basis for planning

experimental validation of SNPs that cause gain or loss of PPIs. Our

implementation is expandable, as datasets used are constantly updated.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiomed.2016.10.003

PMID: 27744178

60. J Struct Funct Genomics. 2016 Dec;17(4):83-99. doi: 10.1007/s10969-016-9208-y.

Epub 2016 Aug 13.

HOMCOS: an updated server to search and model complex 3D structures.

Kawabata T(1).

Author information:

(1)Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka,

565-0871, Japan. kawabata@protein.osaka-u.ac.jp.

The HOMCOS server ( http://homcos.pdbj.org ) was updated for both searching and

modeling the 3D complexes for all molecules in the PDB. As compared to the

previous HOMCOS server, the current server targets all of the molecules in the

PDB including proteins, nucleic acids, small compounds and metal ions. Their

binding relationships are stored in the database. Five services are available for

users. For the services "Modeling a Homo Protein Multimer" and "Modeling a Hetero

Protein Multimer", a user can input one or two proteins as the queries, while for

the service "Protein-Compound Complex", a user can input one chemical compound

and one protein. The server searches similar molecules by BLAST and KCOMBU. Based

on each similar complex found, a simple sequence-replaced model is quickly

generated by replacing the residue names and numbers with those of the query

protein. A target compound is flexibly superimposed onto the template compound

using the program fkcombu. If monomeric 3D structures are input as the query,

then template-based docking can be performed. For the service "Searching Contact

Molecules for a Query Protein", a user inputs one protein sequence as the query,

and then the server searches for its homologous proteins in PDB and summarizes

their contacting molecules as the predicted contacting molecules. The results are

summarized in "Summary Bars" or "Site Table"display. The latter shows the results

as a one-site-one-row table, which is useful for annotating the effects of

mutations. The service "Searching Contact Molecules for a Query Compound" is also

available.

DOI: 10.1007/s10969-016-9208-y

PMCID: PMC5274653

PMID: 27522608

61. Mar Genomics. 2016 Dec;30:67-71. doi: 10.1016/j.margen.2016.10.004. Epub 2016 Oct

11.

Having a BLAST: Searchable transcriptome resources for the gilthead sea bream and

the European sea bass.

Louro B(1), Marques JP(2), Power DM(3), Canário AV(3).

Author information:

(1)Centro de Ciências do Mar, Universidade do Algarve, Campus de Gambelas,

8005-139 Faro, Portugal. Electronic address: blouro@ualg.pt. (2)CIBIO, Centro de

Investigação em Biodiversidade e Recursos Genéticos, InBIO, Laboratório

Associado, Universidade do Porto. Campus Agrário de Vairão, 4485-661 Vairão,

Portugal. (3)Centro de Ciências do Mar, Universidade do Algarve, Campus de

Gambelas, 8005-139 Faro, Portugal.

The gilthead sea bream (Sparus aurata) and the European sea bass (Dicentrarchus

labrax) are the most important aquaculture species in the Mediterranean Sea and

since the last decade it has been seen an exponential increase in their available

molecular resources. In order to improve accessibility to transcriptome

resources, Expressed Sequence Tags (ESTs), mRNA sequences and raw read sequences

were assembled and deposited in BLAST queryable databases. The publicly available

sea bream and sea bass sequences (6.4 and 247.5 million) generated 45,094 and

68,117 assembled sequences, with, respectively, arithmetic mean size of 998 and

2125bp and N50 of 1302 and 2966bp. The assemblies will be regularly updated and

new analytical tools added to the web server at http://sea.ccmar.ualg.pt.

Copyright © 2016 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.margen.2016.10.004

PMID: 27742405

62. Mol Genet Genomics. 2016 Dec;291(6):2225-2229. Epub 2016 Sep 2.

Identifying N (6)-methyladenosine sites in the Arabidopsis thaliana

transcriptome.

Chen W(1), Feng P(2), Ding H(3), Lin H(4).

Author information:

(1)Department of Physics, School of Sciences, and Center for Genomics and

Computational Biology, North China University of Science and Technology,

Tangshan, 063000, China. chenweiimu@gmail.com. (2)School of Public Health, North

China University of Science and Technology, Tangshan, 063000, China. (3)Key

Laboratory for Neuro-Information of Ministry of Education, Center of

Bioinformatics and Center for Information in Biomedicine, School of Life Science

and Technology, University of Electronic Science and Technology of China,

Chengdu, 610054, China. (4)Key Laboratory for Neuro-Information of Ministry of

Education, Center of Bioinformatics and Center for Information in Biomedicine,

School of Life Science and Technology, University of Electronic Science and

Technology of China, Chengdu, 610054, China. hlin@uestc.edu.cn.

N (6)-Methyladenosine (m(6)A) plays important roles in many biological processes.

The knowledge of the distribution of m(6)A is helpful for understanding its

regulatory roles. Although the experimental methods have been proposed to detect

m(6)A, the resolutions of these methods are still unsatisfying especially for

Arabidopsis thaliana. Benefitting from the experimental data, in the current

work, a support vector machine-based method was proposed to identify m(6)A sites

in A. thaliana transcriptome. The proposed method was validated on a benchmark

dataset using jackknife test and was also validated by identifying

strain-specific m(6)A sites in A. thaliana. The obtained predictive results

indicate that the proposed method is quite promising. For the convenience of

experimental biologists, an online webserver for the proposed method was built,

which is freely available at http://lin.uestc.edu.cn/server/M6ATH . These results

indicate that the proposed method holds a potential to become an elegant tool in

identifying m(6)A site in A. thaliana.

DOI: 10.1007/s00438-016-1243-7

PMID: 27590733 [Indexed for MEDLINE]

63. Nat Protoc. 2016 Dec;11(12):2529-2548. doi: 10.1038/nprot.2016.150. Epub 2016 Nov

17.

Indel variant analysis of short-read sequencing data with Scalpel.

Fang H(1,)(2,)(3), Bergmann EA(4), Arora K(4), Vacic V(4), Zody MC(4), Iossifov

I(1), O'Rawe JA(2,)(3), Wu Y(2,)(3), Jimenez Barron LT(2,)(5), Rosenbaum J(1),

Ronemus M(1), Lee YH(1), Wang Z(1), Dikoglu E(2), Jobanputra V(2,)(6), Lyon

GJ(2,)(3), Wigler M(1), Schatz MC(1,)(7), Narzisi G(1,)(4).

Author information:

(1)Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold

Spring Harbor, New York, USA. (2)Stanley Institute for Cognitive Genomics, Cold

Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. (3)Stony Brook

University, Stony Brook, New York, USA. (4)New York Genome Center, New York, New

York, USA. (5)Centro de Ciencias Genomicas, Universidad Nacional Autonoma de

Mexico, Cuernavaca, Mexico. (6)Columbia University Medical Center, New York, New

York, USA. (7)Department of Computer Science, Johns Hopkins University,

Baltimore, Maryland, USA.

As the second most common type of variation in the human genome, insertions and

deletions (indels) have been linked to many diseases, but the discovery of indels

of more than a few bases in size from short-read sequencing data remains

challenging. Scalpel (http://scalpel.sourceforge.net) is an open-source software

for reliable indel detection based on the microassembly technique. It has been

successfully used to discover mutations in novel candidate genes for autism, and

it is extensively used in other large-scale studies of human diseases. This

protocol gives an overview of the algorithm and describes how to use Scalpel to

perform highly accurate indel calling from whole-genome and whole-exome

sequencing data. We provide detailed instructions for an exemplary family-based

de novo study, but we also characterize the other two supported modes of

operation: single-sample and somatic analysis. Indel normalization, visualization

and annotation of the mutations are also illustrated. Using a standard server,

indel discovery and characterization in the exonic regions of the example

sequencing data can be completed in ∼5 h after read mapping.

DOI: 10.1038/nprot.2016.150

PMID: 27854363

64. PLoS One. 2016 Dec 1;11(12):e0167345. doi: 10.1371/journal.pone.0167345.

eCollection 2016.

DNABP: Identification of DNA-Binding Proteins Based on Feature Selection Using a

Random Forest and Predicting Binding Residues.

Ma X(1), Guo J(2), Sun X(2).

Author information:

(1)School of Science, Nanjing Audit University, Nanjing, China. (2)State Key

Laboratory of Bioelectronics, School of Biological Science and Medical

Engineering, Southeast University, Nanjing, China.

DNA-binding proteins are fundamentally important in cellular processes. Several

computational-based methods have been developed to improve the prediction of

DNA-binding proteins in previous years. However, insufficient work has been done

on the prediction of DNA-binding proteins from protein sequence information. In

this paper, a novel predictor, DNABP (DNA-binding proteins), was designed to

predict DNA-binding proteins using the random forest (RF) classifier with a

hybrid feature. The hybrid feature contains two types of novel sequence features,

which reflect information about the conservation of physicochemical properties of

the amino acids, and the binding propensity of DNA-binding residues and

non-binding propensities of non-binding residues. The comparisons with each

feature demonstrated that these two novel features contributed most to the

improvement in predictive ability. Furthermore, to improve the prediction

performance of the DNABP model, feature selection using the minimum redundancy

maximum relevance (mRMR) method combined with incremental feature selection (IFS)

was carried out during the model construction. The results showed that the DNABP

model could achieve 86.90% accuracy, 83.76% sensitivity, 90.03% specificity and a

Matthews correlation coefficient of 0.727. High prediction accuracy and

performance comparisons with previous research suggested that DNABP could be a

useful approach to identify DNA-binding proteins from sequence information. The

DNABP web server system is freely available at http://www.cbi.seu.edu.cn/DNABP/.

DOI: 10.1371/journal.pone.0167345

PMCID: PMC5132331

PMID: 27907159

Conflict of interest statement: The authors have declared that no competing

interests exist.

65. Schweiz Arch Tierheilkd. 2016 Dec;158(12):805-810.

[AntibioticScout: Online tool for antimicrobial stewardship in veterinary

medicine].

[Article in German; Abstract available in German from the publisher]

Peter R(1), Müntener C(1), Demuth D(1), Heim D(2), Mevissen M(3),

Schüpbach-Regula G(4), Schuller S(5), Stucki F(2), Willi B(6), Naegeli H(1).

Author information:

(1)Institut für Veterinärpharmakologie und -toxikologie der Universität Zürich.

(2)Bundesamt für Lebensmittelsicherheit und Veterinärwesen BLV, Bern.

(3)Abteilung für Veteinärpharmakologie und -toxikologie der Universität Bern.

(4)Veterinary Public Health Institute (VPHI), Universität Bern. (5)Klinik für

Kleintiermedizin der Universität Bern. (6)Klinik für Kleintiermedizin der

Universität Zürich.

INTRODUCTION: Resistances to antimicrobials pose serious public health

challenges. This issue concerns both human and veterinary medicine and can only

be solved by a multidisciplinary approach. A comprehensive concept is, therefore,

being worked out within the StAR (strategy antibiotic resistance) program in

order to preserve the effectiveness of antibiotics for humans as well as animals.

In this context, the AntibioticScout (www.AntibioticScout. ch) offers a new

online tool for the prudent use of antibiotics in veterinary medicine. By

involving all stakeholders, the guidelines included in the AntibioticScout will

result in a nationwide accepted standard for the treatment of bacterial

infections in animals. An additional system for the rapid reporting of cases of

suspected lack of efficacy of antimicrobials is integrated to allow early

detection of emerging resistance and the immediate launch of risk mitigation

measures. A first version of the AntibioticScout for the treatment of dogs, cats

and horses is available by the end of 2016. All stakeholders are now invited to

contribute to the development of the AntibioticScout decision support.

Publisher: Resistenzen gegenüber antimikrobiellen Wirkstoffen stellen das

Gesundheitssystem vor grosse Herausforderungen. Dieses Problem betrifft die

Human- wie die Veterinärmedizin und kann nur bereichsübergreifend gelöst werden.

Deshalb wird mit dem StAR (Strategie Antibiotikaresistenzen)-Programm ein

umfassendes Gesamtkonzept ausgearbeitet, um die Wirksamkeit von Antibiotika für

Mensch und Tier langfristig zu erhalten. In diesem Zusammenhang bieten wir mit

der Entwicklung des AntibioticScout (www.AntibioticScout.ch) ein neues

Online-Instrument an, um den verantwortungsvollen Einsatz von Antibiotika in der

Veterinärmedizin zu unterstützen. Unter Einbezug aller Interessenvertreter sollen

Leitlinien zu einem national akzeptierten Standard zur Therapie von bakteriellen

Infektionen bei Tieren erarbeitet und im AntibioticScout aufgeführt werden.

Darüber hinaus wird ein Sofortsystem zur Meldung von Unwirksamkeiten angeboten,

um Risiken frühzeitig zu erkennen und entsprechende Massnahmen zu ergreifen. Eine

erste Fassung des AntibioticScout für die Behandlung von Hunden, Katzen und

Pferden ist ab Ende 2016 verfügbar. Alle Interessenvertreter sind nun eingeladen,

sich am Aufbau des AntibioticScout zu beteiligen.Publisher: Les résistances face

aux substances antimicrobiennes placent le système de santé face à de grands

défis. Ce problème touche aussi bien la médecine humaine que vétérinaire et ne

peut être réglé que de façon transversale. C’est pour cette raison qu’a été

développé, avec le programme StAR (Strategie Antibiotikaresistenzen), un concept

global pour assurer à long terme l’efficacité des antibiotiques aussi bien chez

les hommes que chez les animaux. Dans ce contexte, nous mettons à disposition,

avec le développement d’AntibioticScout (www.AntibioticScout. ch), un nouvel

outil en ligne pour soutenir un usage responsable des antibiotiques en médecine

vétérinaire. Avec le concours de représentants de tous les milieux intéressés, il

s’agit de développer les lignes directrices d’un standard accepté au plan

national pour le traitement des infections bactériennes chez les animaux et de le

mettre à disposition dans AntibioticScout. En outre un système d’annonce

immédiate en cas d’inefficacité sera mis en place, afin de repérer précocement

les risques et de prendre les mesures correspondantes. Une première version

d’AntibioticScout pour le traitement des chiens, chats et chevaux est disponible

dès fin 2016. Tous les cercles intéressés sont dès maintenant invités à

contribuer au développement d’Antibiotic Scout.Publisher: Il sistema sanitario è

messo a dura prova dalla resistenza agli agenti antimicrobici. Questo problema

riguarda sia la medicina umana che quella veterinaria e può essere risolto solo

in modo trasversale. Pertanto, è stato redatto sotto il nome di programma StAR

(strategia contro la resistenza agli antibiotici) un approccio integrato per

preservare l’efficacia degli antibiotici per l’uomo e per gli animali sul lungo

termine. In questo contesto, con lo sviluppo di AntibioticScout

(www.AntibioticScout. ch) proponiamo un nuovo strumento online per sostenere

l’uso responsabile degli antibiotici in medicina veterinaria. Con il sostegno di

tutte le parti interessate saranno sviluppate delle linee guida, con uno standard

riconosciuto a livello nazionale, per il trattamento delle infezioni batteriche

negli animali. Queste saranno inserite in AntibioticScout. Inoltre, alfine di

identificare in anticipo i rischi e di prendere le misure necessarie, verrà

offerto un sistema istantaneo per riportare le inefficienze. Una versione

iniziale di AntibioticScout per il trattamento di cani, gatti e cavalli sarà

disponibile a partire dalla fine del 2016. Tutte le parti interessate sono

invitate a partecipare alla realizzazione di Antibiotic- Scout.

DOI: 10.17236/sat00095

PMID: 27934622

66. PLoS One. 2016 Nov 28;11(11):e0166965. doi: 10.1371/journal.pone.0166965.

eCollection 2016.

Evolutionary Algorithm for RNA Secondary Structure Prediction Based on Simulated

SHAPE Data.

Montaseri S(1), Ganjtabesh M(1), Zare-Mirakabad F(2).

Author information:

(1)Department of Computer Science, School of Mathematics, Statistics, and

Computer Science, University of Tehran, Tehran, Iran. (2)Department of Computer

Science, Faculty of Mathematics and Computer Science, Amirkabir University of

Technology, Tehran, Iran.

BACKGROUND: Non-coding RNAs perform a wide range of functions inside the living

cells that are related to their structures. Several algorithms have been proposed

to predict RNA secondary structure based on minimum free energy. Low prediction

accuracy of these algorithms indicates that free energy alone is not sufficient

to predict the functional secondary structure. Recently, the obtained information

from the SHAPE experiment greatly improves the accuracy of RNA secondary

structure prediction by adding this information to the thermodynamic free energy

as pseudo-free energy.

METHOD: In this paper, a new method is proposed to predict RNA secondary

structure based on both free energy and SHAPE pseudo-free energy. For each RNA

sequence, a population of secondary structures is constructed and their SHAPE

data are simulated. Then, an evolutionary algorithm is used to improve each

structure based on both free and pseudo-free energies. Finally, a structure with

minimum summation of free and pseudo-free energies is considered as the predicted

RNA secondary structure.

RESULTS AND CONCLUSIONS: Computationally simulating the SHAPE data for a given

RNA sequence requires its secondary structure. Here, we overcome this limitation

by employing a population of secondary structures. This helps us to simulate the

SHAPE data for any RNA sequence and consequently improves the accuracy of RNA

secondary structure prediction as it is confirmed by our experiments. The source

code and web server of our proposed method are freely available at

http://mostafa.ut.ac.ir/ESD-Fold/.

DOI: 10.1371/journal.pone.0166965

PMCID: PMC5125645

PMID: 27893832

Conflict of interest statement: The authors have declared that no competing

interests exist.

67. BMC Genomics. 2016 Nov 18;17(1):938.

The ChIP-Seq tools and web server: a resource for analyzing ChIP-seq and other

types of genomic data.

Ambrosini G(1,)(2), Dreos R(1,)(2), Kumar S(1,)(2), Bucher P(3,)(4).

Author information:

(1)School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL),

CH-1015, Lausanne, Switzerland. (2)Swiss Institute of Bioinformatics (SIB),

CH-1015, Lausanne, Switzerland. (3)School of Life Sciences, Ecole Polytechnique

Fédérale de Lausanne (EPFL), CH-1015, Lausanne, Switzerland.

philipp.bucher@epfl.ch. (4)Swiss Institute of Bioinformatics (SIB), CH-1015,

Lausanne, Switzerland. philipp.bucher@epfl.ch.

BACKGROUND: ChIP-seq and related high-throughput chromatin profilig assays

generate ever increasing volumes of highly valuable biological data. To make

sense out of it, biologists need versatile, efficient and user-friendly tools for

access, visualization and itegrative analysis of such data.

RESULTS: Here we present the ChIP-Seq command line tools and web server,

implementing basic algorithms for ChIP-seq data analysis starting with a read

alignment file. The tools are optimized for memory-efficiency and speed thus

allowing for processing of large data volumes on inexpensive hardware. The web

interface provides access to a large database of public data. The ChIP-Seq tools

have a modular and interoperable design in that the output from one application

can serve as input to another one. Complex and innovative tasks can thus be

achieved by running several tools in a cascade.

CONCLUSIONS: The various ChIP-Seq command line tools and web services either

complement or compare favorably to related bioinformatics resources in terms of

computational efficiency, ease of access to public data and interoperability with

other web-based tools. The ChIP-Seq server is accessible at

http://ccg.vital-it.ch/chipseq/ .

DOI: 10.1186/s12864-016-3288-8

PMCID: PMC5116162

PMID: 27863463

68. BMC Genomics. 2016 Nov 16;17(1):931.

AnnoLnc: a web server for systematically annotating novel human lncRNAs.

Hou M(1), Tang X(1,)(2), Tian F(1,)(3), Shi F(1), Liu F(1), Gao G(4).

Author information:

(1)State Key Laboratory of Protein and Plant Gene Research, College of Life

Sciences, Center for Bioinformatics, Peking University, Beijing, 100871, P.R.

China. (2)Present address: Department of Hematology, St. Jude Children's Research

Hospital, Memphis, TN, USA. (3)Peking-Tsinghua Center for Life Sciences, Academy

for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, P.R.

China. (4)State Key Laboratory of Protein and Plant Gene Research, College of

Life Sciences, Center for Bioinformatics, Peking University, Beijing, 100871,

P.R. China. gaog@mail.cbi.pku.edu.cn.

BACKGROUND: Long noncoding RNAs (lncRNAs) have been shown to play essential roles

in almost every important biological process through multiple mechanisms.

Although the repertoire of human lncRNAs has rapidly expanded, their biological

function and regulation remain largely elusive, calling for a systematic and

integrative annotation tool.

RESULTS: Here we present AnnoLnc ( http://annolnc.cbi.pku.edu.cn ), a one-stop

portal for systematically annotating novel human lncRNAs. Based on more than 700

data sources and various tool chains, AnnoLnc enables a systematic annotation

covering genomic location, secondary structure, expression patterns,

transcriptional regulation, miRNA interaction, protein interaction, genetic

association and evolution. An intuitive web interface is available for

interactive analysis through both desktops and mobile devices, and programmers

can further integrate AnnoLnc into their pipeline through standard JSON-based Web

Service APIs.

CONCLUSIONS: To the best of our knowledge, AnnoLnc is the only web server to

provide on-the-fly and systematic annotation for newly identified human lncRNAs.

Compared with similar tools, the annotation generated by AnnoLnc covers a much

wider spectrum with intuitive visualization. Case studies demonstrate the power

of AnnoLnc in not only rediscovering known functions of human lncRNAs but also

inspiring novel hypotheses.

DOI: 10.1186/s12864-016-3287-9

PMCID: PMC5112684

PMID: 27852242

69. Bioinformatics. 2016 Nov 15;32(22):3489-3491. Epub 2016 Aug 2.

NET-GE: a web-server for NETwork-based human gene enrichment.

Bovo S(1), Di Lena P(2), Martelli PL(1), Fariselli P(3), Casadio R(1,)(4).

Author information:

(1)Biocomputing Group, CIG, Interdepartmental Center «Luigi Galvani» for

Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of

Bologna, Bologna, Italy. (2)DISI, University of Bologna, Bologna, Italy. (3)BCA,

University of Padova, Padova, Italy. (4)Interdepartmental Center «Giorgio Prodi»

for Cancer Research, University of Bologna, Bologna, Italy.

MOTIVATION: Gene enrichment is a requisite for the interpretation of biological

complexity related to specific molecular pathways and biological processes.

Furthermore, when interpreting NGS data and human variations, including those

related to pathologies, gene enrichment allows the inclusion of other genes that

in the human interactome space may also play important key roles in the emergency

of the phenotype. Here, we describe NET-GE, a web server for associating

biological processes and pathways to sets of human proteins involved in the same

phenotype RESULTS: NET-GE is based on protein-protein interaction networks,

following the notion that for a set of proteins, the context of their specific

interactions can better define their function and the processes they can be

related to in the biological complexity of the cell. Our method is suited to

extract statistically validated enriched terms from Gene Ontology, KEGG and

REACTOME annotation databases. Furthermore, NET-GE is effective even when the

number of input proteins is small.

AVAILABILITY AND IMPLEMENTATION: NET-GE web server is publicly available and

accessible at http://net-ge.biocomp.unibo.it/enrich CONTACT:

gigi@biocomp.unibo.itSupplementary information: Supplementary data are available

at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw508

PMID: 27485441

70. PLoS One. 2016 Nov 11;11(11):e0166126. doi: 10.1371/journal.pone.0166126.

eCollection 2016.

Vidjil: A Web Platform for Analysis of High-Throughput Repertoire Sequencing.

Duez M(1,)(2), Giraud M(3,)(4), Herbert R(3,)(4), Rocher T(3,)(4), Salson

M(3,)(4), Thonier F(5).

Author information:

(1)School of Social and Community Medicine, University of Bristol, Bristol,

United Kingdom. (2)SIRIC ONCOLille, 59000 Lille, France. (3)Université de Lille,

CNRS, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et

Automatique de Lille, 59000 Lille, France. (4)Inria Lille, 59650 Villeneuve

d'Ascq, France. (5)Inserm UMR1151, Institut Necker-Enfants Malades, Paris

Descartes University and Necker–Enfants Malades Hospital–Paris, France.

Erratum in

PLoS One. 2017 Feb 9;12 (2):e0172249.

BACKGROUND: The B and T lymphocytes are white blood cells playing a key role in

the adaptive immunity. A part of their DNA, called the V(D)J recombinations, is

specific to each lymphocyte, and enables recognition of specific antigenes.

Today, with new sequencing techniques, one can get billions of DNA sequences from

these regions. With dedicated Repertoire Sequencing (RepSeq) methods, it is now

possible to picture population of lymphocytes, and to monitor more accurately the

immune response as well as pathologies such as leukemia.

METHODS AND RESULTS: Vidjil is an open-source platform for the interactive

analysis of high-throughput sequencing data from lymphocyte recombinations. It

contains an algorithm gathering reads into clonotypes according to their V(D)J

junctions, a web application made of a sample, experiment and patient database

and a visualization for the analysis of clonotypes along the time. Vidjil is

implemented in C++, Python and Javascript and licensed under the GPLv3

open-source license. Source code, binaries and a public web server are available

at http://www.vidjil.org and at http://bioinfo.lille.inria.fr/vidjil. Using the

Vidjil web application consists of four steps: 1. uploading a raw sequence file

(typically a FASTQ); 2. running RepSeq analysis software; 3. visualizing the

results; 4. annotating the results and saving them for future use. For the

end-user, the Vidjil web application needs no specific installation and just

requires a connection and a modern web browser. Vidjil is used by labs in

hematology or immunology for research and clinical applications.

DOI: 10.1371/journal.pone.0166126

PMCID: PMC5106020

PMID: 27835690

Conflict of interest statement: The authors have declared that no competing

interests exist.

71. J Cheminform. 2016 Nov 4;8:61. eCollection 2016.

ClassyFire: automated chemical classification with a comprehensive, computable

taxonomy.

Djoumbou Feunang Y(1), Eisner R(2), Knox C(3), Chepelev L(4), Hastings J(5), Owen

G(5), Fahy E(6), Steinbeck C(5), Subramanian S(6), Bolton E(7), Greiner R(8),

Wishart DS(9).

Author information:

(1)Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E8

Canada. (2)Jobber - Field Service Software, 10520 Jasper Ave, Edmonton, AB T5J

1Z7 Canada. (3)Department of Computing Science, University of Alberta, Edmonton,

AB T6G 2E8 Canada. (4)Department of Medical Imaging, The Ottawa Hospital,

University of Ottawa, Civic Campus, 1053 Carling Ave, Ottawa, ON K1Y 4E9 Canada.

(5)European Molecular Biology Laboratory - European Bioinformatics Institute

(EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge UK. (6)Department of

Bioengineering, University of California, La Jolla, San Diego, CA 92093 USA.

(7)Department of Health and Human Services, National Center for Biotechnology

Information, National Library of Medicine, National Institutes of Health, 8600

Rockville Pike, Bethesda, MD 20894 USA. (8)Department of Computing Science,

University of Alberta, Edmonton, AB T6G 2E8 Canada ; Department of Computing

Science 2-21 Athabasca Hall, Alberta Innovates Centre for Machine Learning

(AICML), University of Alberta, Edmonton, AB T6G 2E8 Canada. (9)Department of

Biological Sciences, University of Alberta, Edmonton, AB T6G 2E8 Canada ;

Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8

Canada ; National Research Council, National Institute for Nanotechnology (NINT),

Edmonton, AB T6G 2M9 Canada ; The Metabolomics Innovation Center, University of

Alberta, Edmonton, AB T6G 2E9 Canada.

BACKGROUND: Scientists have long been driven by the desire to describe, organize,

classify, and compare objects using taxonomies and/or ontologies. In contrast to

biology, geology, and many other scientific disciplines, the world of chemistry

still lacks a standardized chemical ontology or taxonomy. Several attempts at

chemical classification have been made; but they have mostly been limited to

either manual, or semi-automated proof-of-principle applications. This is

regrettable as comprehensive chemical classification and description tools could

not only improve our understanding of chemistry but also improve the linkage

between chemistry and many other fields. For instance, the chemical

classification of a compound could help predict its metabolic fate in humans, its

druggability or potential hazards associated with it, among others. However, the

sheer number (tens of millions of compounds) and complexity of chemical

structures is such that any manual classification effort would prove to be near

impossible.

RESULTS: We have developed a comprehensive, flexible, and computable, purely

structure-based chemical taxonomy (ChemOnt), along with a computer program

(ClassyFire) that uses only chemical structures and structural features to

automatically assign all known chemical compounds to a taxonomy consisting of

>4800 different categories. This new chemical taxonomy consists of up to 11

different levels (Kingdom, SuperClass, Class, SubClass, etc.) with each of the

categories defined by unambiguous, computable structural rules. Furthermore each

category is named using a consensus-based nomenclature and described (in English)

based on the characteristic common structural properties of the compounds it

contains. The ClassyFire webserver is freely accessible at

http://classyfire.wishartlab.com/. Moreover, a Ruby API version is available at

https://bitbucket.org/wishartlab/classyfire\_api, which provides programmatic

access to the ClassyFire server and database. ClassyFire has been used to

annotate over 77 million compounds and has already been integrated into other

software packages to automatically generate textual descriptions for, and/or

infer biological properties of over 100,000 compounds. Additional examples and

applications are provided in this paper.

CONCLUSION: ClassyFire, in combination with ChemOnt (ClassyFire's comprehensive

chemical taxonomy), now allows chemists and cheminformaticians to perform

large-scale, rapid and automated chemical classification. Moreover, a freely

accessible API allows easy access to more than 77 million "ClassyFire" classified

compounds. The results can be used to help annotate well studied, as well as

lesser-known compounds. In addition, these chemical classifications can be used

as input for data integration, and many other cheminformatics-related tasks.

DOI: 10.1186/s13321-016-0174-y

PMCID: PMC5096306

PMID: 27867422

72. Sci Rep. 2016 Nov 2;6:35996. doi: 10.1038/srep35996.

DPDR-CPI, a server that predicts Drug Positioning and Drug Repositioning via

Chemical-Protein Interactome.

Luo H(1), Zhang P(2), Cao XH(3), Du D(1), Ye H(1), Huang H(1), Li C(1), Qin S(1),

Wan C(1), Shi L(4), He L(1), Yang L(1).

Author information:

(1)Bio-X Institutes, Shanghai Jiao Tong University, Shanghai 200030, China.

(2)Center for Computational Health, IBM T.J. Watson Research Center, Yorktown

Heights, NY 10598, USA. (3)Center for Data Analytics and Biomedical Informatics,

Temple University, Philadelphia, PA 19122, USA. (4)Collaborative Innovation

Center for Genetics and Development, State Key Laboratory of Genetic Engineering

and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences,

Fudan University, Shanghai 200438, China.

The cost of developing a new drug has increased sharply over the past years. To

ensure a reasonable return-on-investment, it is useful for drug discovery

researchers in both industry and academia to identify all the possible

indications for early pipeline molecules. For the first time, we propose the term

computational "drug candidate positioning" or "drug positioning", to describe the

above process. It is distinct from drug repositioning, which identifies new uses

for existing drugs and maximizes their value. Since many therapeutic effects are

mediated by unexpected drug-protein interactions, it is reasonable to analyze the

chemical-protein interactome (CPI) profiles to predict indications. Here we

introduce the server DPDR-CPI, which can make real-time predictions based only on

the structure of the small molecule. When a user submits a molecule, the server

will dock it across 611 human proteins, generating a CPI profile of features that

can be used for predictions. It can suggest the likelihood of relevance of the

input molecule towards ~1,000 human diseases with top predictions listed.

DPDR-CPI achieved an overall AUROC of 0.78 during 10-fold cross-validations and

AUROC of 0.76 for the independent validation. The server is freely accessible via

http://cpi.bio-x.cn/dpdr/.

DOI: 10.1038/srep35996

PMCID: PMC5090963

PMID: 27805045

Conflict of interest statement: Lun Yang is a current employee at Bayer Pharma AG

and Heng Luo is a current postdoctoral researcher at IBM. However, this study was

based on their previous work at Shanghai Jiao Tong University and Bayer Pharma AG

was not involved.

73. Anal Chem. 2016 Nov 1;88(21):10395-10403. Epub 2016 Oct 13.

Web Server for Peak Detection, Baseline Correction, and Alignment in

Two-Dimensional Gas Chromatography Mass Spectrometry-Based Metabolomics Data.

Tian TF(1,)(2), Wang SY(1,)(2), Kuo TC(2,)(3), Tan CE(1,)(2), Chen GY(2,)(4), Kuo

CH(2,)(4,)(5), Chen CS(6), Chan CC(6), Lin OA(3), Tseng YJ(1,)(2,)(4,)(3).

Author information:

(1)Department of Computer Science and Information Engineering, National Taiwan

University , No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan. (2)The

Metabolomics Core Laboratory, Center of Genomic Medicine, National Taiwan

University , No. 2, Syu-Jhou Road, Taipei 10055, Taiwan. (3)Graduate Institute of

Biomedical Electronics and Bioinformatics, National Taiwan University , No. 1,

Sec. 4, Roosevelt Road, Taipei 10617, Taiwan. (4)School of Pharmacy, College of

Medicine, National Taiwan University , No. 33, Linsen S. Road, Taipei 100,

Taiwan. (5)Department of Pharmacy, National Taiwan University Hospital, National

Taiwan University , No. 1, Changde Street, Taipei 10048, Taiwan. (6)Institute of

Occupational Medicine and Industrial Hygiene, National Taiwan University , No.

17, Xuzhou Road, Taipei 100, Taiwan.

Two-dimensional gas chromatography time-of-flight mass spectrometry

(GC×GC/TOF-MS) is superior for chromatographic separation and provides great

sensitivity for complex biological fluid analysis in metabolomics. However,

GC×GC/TOF-MS data processing is currently limited to vendor software and

typically requires several preprocessing steps. In this work, we implement a

web-based platform, which we call GC(2)MS, to facilitate the application of

recent advances in GC×GC/TOF-MS, especially for metabolomics studies. The core

processing workflow of GC(2)MS consists of blob/peak detection, baseline

correction, and blob alignment. GC(2)MS treats GC×GC/TOF-MS data as pictures and

clusters the pixels as blobs according to the brightness of each pixel to

generate a blob table. GC(2)MS then aligns the blobs of two GC×GC/TOF-MS data

sets according to their distance and similarity. The blob distance and similarity

are the Euclidean distance of the first and second retention times of two blobs

and the Pearson's correlation coefficient of the two mass spectra, respectively.

GC(2)MS also directly corrects the raw data baseline. The analytical performance

of GC(2)MS was evaluated using GC×GC/TOF-MS data sets of Angelica sinensis

compounds acquired under different experimental conditions and of human plasma

samples. The results show that GC(2)MS is an easy-to-use tool for detecting peaks

and correcting baselines, and GC(2)MS is able to align GC×GC/TOF-MS data sets

acquired under different experimental conditions. GC(2)MS is freely accessible at

http://gc2ms.web.cmdm.tw .

DOI: 10.1021/acs.analchem.6b00755

PMID: 27673369

74. Bioinformatics. 2016 Nov 1;32(21):3330-3332. Epub 2016 Jul 4.

The SMAL web server: global multiple network alignment from pairwise alignments.

Dohrmann J(1), Singh R(2).

Author information:

(1)Department of Computer Science, San Francisco State University, San Francisco,

CA 94132, USA. (2)Department of Computer Science, San Francisco State University,

San Francisco, CA 94132, USA Center for Discovery and Innovation in Parasitic

Diseases, University of California, San Diego, CA 92093, USA.

MOTIVATION: Alignments of protein-protein interaction networks (PPIN) can be used

to predict protein function, study conserved aspects of the interactome, and to

establish evolutionary correspondences. Within this problem context, determining

multiple network alignments (MNA) is a significant challenge that involves high

computational complexity. A limited number of public MNA implementations are

available currently and the majority of the pairwise network alignment (PNA)

algorithms do not have MNA counterparts. Furthermore, current MNA algorithms do

not allow choosing a specific PPIN relative to which an MNA could be constructed.

Also, once an MNA is obtained, it cannot easily be modified, such as through

addition of a new network, without expensive re-computation of the entire MNA.

RESULTS: SMAL (Scaffold-Based Multiple Network Aligner) is a public, open-source,

web-based application for determining MNAs from existing PNAs that addresses all

the aforementioned challenges. With SMAL, PNAs can be combined rapidly to obtain

an MNA. The software also supports visualization and user-data interactions to

facilitate exploratory analysis and sensemaking. SMAL is especially useful when

multiple alignments relative to a particular PPIN are required; furthermore, SMAL

alignments are persistent in that existing correspondences between networks

(obtained during PNA or MNA) are not lost as new networks are added. In

comparative studies alongside existent MNA techniques, SMAL MNAs were found to be

superior per a number of measures, such as the total number of identified

homologs and interologs as well as the fraction of all identified correspondences

that are functionally similar or homologous to the scaffold. While directed

primarily at PPIN-alignment, SMAL is a generic network aligner and may be applied

to arbitrary networks.Availability information: The SMAL web server and source

code is available at: http://haddock6.sfsu.edu/smal/ CONTACT:

rahul@sfsu.eduSupplementary information: Supplementary data are available at

Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw402

PMID: 27378297

75. Bioinformatics. 2016 Nov 1;32(21):3339-3341. Epub 2016 Jul 4.

MetaPred2CS: a sequence-based meta-predictor for protein-protein interactions of

prokaryotic two-component system proteins.

Kara A(1), Vickers M(1), Swain M(1), Whitworth DE(1), Fernandez-Fuentes N(1).

Author information:

(1)Institute of Biological, Environmental and Rural Sciences, Aberystwyth

University, Aberystwyth SY23 3EB, UK.

MOTIVATION: Two-component systems (TCS) are the main signalling pathways of

prokaryotes, and control a wide range of biological phenomena. Their functioning

depends on interactions between TCS proteins, the specificity of which is poorly

understood.

RESULTS: The MetaPred2CS web-server interfaces a sequence-based meta-predictor

specifically designed to predict pairing of the histidine kinase and

response-regulator proteins forming TCSs. MetaPred2CS integrates six

sequence-based methods using a support vector machine classifier and has been

intensively tested under different benchmarking conditions: (i) species specific

gene sets; (ii) neighbouring versus orphan pairs; and (iii) k-fold cross

validation on experimentally validated datasets.

AVAILABILITY AND IMPLEMENTATION: Web server at:

http://metapred2cs.ibers.aber.ac.uk/, Source code:

https://github.com/martinjvickers/MetaPred2CS or implemented as Virtual Machine

at: http://metapred2cs.ibers.aber.ac.uk/download CONTACT:

naf4@aber.ac.ukSupplementary information: Supplementary data are available at

Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw403

PMID: 27378293

76. Bioinformatics. 2016 Nov 1;32(21):3333-3335. Epub 2016 Jul 4.

w4CSeq: software and web application to analyze 4C-seq data.

Cai M(1), Gao F(2), Lu W(3), Wang K(4).

Author information:

(1)Department of Preventive Medicine Eli and Edythe Broad Center for Regenerative

Medicine and Stem Cell Research Zilkha Neurogenetic Institute. (2)Eli and Edythe

Broad Center for Regenerative Medicine and Stem Cell Research Zilkha Neurogenetic

Institute. (3)Eli and Edythe Broad Center for Regenerative Medicine and Stem Cell

Research Department of Stem Cell Biology and Regenerative Medicine. (4)Zilkha

Neurogenetic Institute Department of Psychiatry, Keck School of Medicine,

University of Southern California, Los Angeles, CA 90033, USA.

Circularized Chromosome Conformation Capture followed by deep sequencing (4C-Seq)

is a powerful technique to identify genome-wide partners interacting with a

pre-specified genomic locus. Here, we present a computational and statistical

approach to analyze 4C-Seq data generated from both enzyme digestion and

sonication fragmentation-based methods. We implemented a command line software

tool and a web interface called w4CSeq, which takes in the raw 4C sequencing data

(FASTQ files) as input, performs automated statistical analysis and presents

results in a user-friendly manner. Besides providing users with the list of

candidate interacting sites/regions, w4CSeq generates figures showing genome-wide

distribution of interacting regions, and sketches the enrichment of key features

such as TSSs, TTSs, CpG sites and DNA replication timing around 4C

sites.AVAILABILITY AND IMPLEMENTATION: Users can establish their own web server

by downloading source codes at https://github.com/WGLab/w4CSeq Additionally, a

demo web server is available at http://w4cseq.wglab.org CONTACT: kaiwang@usc.edu

or wangelu@usc.eduSupplementary information: Supplementary data are available at

Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw408

PMCID: PMC5079477 [Available on 2017-11-01]

PMID: 27378289

77. Bioinformatics. 2016 Nov 1;32(21):3342-3344. Epub 2016 Jun 29.

Accounting for pairwise distance restraints in FFT-based protein-protein docking.

Xia B(1), Vajda S(2), Kozakov D(3).

Author information:

(1)Department of Biomedical Engineering. (2)Department of Biomedical Engineering

Department of Chemistry, Boston University, Boston, MA 02215, USA. (3)Department

of Applied Mathematics and Statistics Laufer Center for Physical and Quantitative

Biology, Stony Brook University, Stony Brook, NY 11794, USA.

ClusPro is a heavily used protein-protein docking server based on the fast

Fourier transform (FFT) correlation approach. While FFT enables global docking,

accounting for pairwise distance restraints using penalty terms in the scoring

function is computationally expensive. We use a different approach and directly

select low energy solutions that also satisfy the given restraints. As expected,

accounting for restraints generally improves the rank of near native predictions,

while retaining or even improving the numerical efficiency of FFT based

docking.AVAILABILITY AND IMPLEMENTATION: The software is freely available as part

of the ClusPro web-based server at http://cluspro.org/nousername.php CONTACT:

midas@laufercenter.org or vajda@bu.eduSupplementary information: Supplementary

data are available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw306

PMID: 27357172

78. Plant J. 2016 Nov;88(3):490-504. doi: 10.1111/tpj.13261. Epub 2016 Oct 5.

New BAR tools for mining expression data and exploring Cis-elements in

Arabidopsis thaliana.

Austin RS(1), Hiu S(1), Waese J(1), Ierullo M(1), Pasha A(1), Wang TT(1), Fan

J(1), Foong C(1), Breit R(1), Desveaux D(1), Moses A(1), Provart NJ(1).

Author information:

(1)Department of Cell & Systems Biology/Centre for the Analysis of Genome

Evolution and Function, University of Toronto, Toronto, ON, M5S 3B2, Canada.

Identifying sets of genes that are specifically expressed in certain tissues or

in response to an environmental stimulus is useful for designing reporter

constructs, generating gene expression markers, or for understanding gene

regulatory networks. We have developed an easy-to-use online tool for defining a

desired expression profile (a modification of our Expression Angler program),

which can then be used to identify genes exhibiting patterns of expression that

match this profile as closely as possible. Further, we have developed another

online tool, Cistome, for predicting or exploring cis-elements in the promoters

of sets of co-expressed genes identified by such a method, or by other methods.

We present two use cases for these tools, which are freely available on the

Bio-Analytic Resource at http://BAR.utoronto.ca.

© 2016 The Authors The Plant Journal © 2016 John Wiley & Sons Ltd.

DOI: 10.1111/tpj.13261

PMID: 27401965

79. RNA Biol. 2016 Nov;13(11):1144-1151. Epub 2016 Sep 7.

SMEpred workbench: A web server for predicting efficacy of chemicallymodified

siRNAs.

Dar SA(1), Gupta AK(1), Thakur A(1), Kumar M(1).

Author information:

(1)a Bioinformatics Center, Institute of Microbial Technology, Council of

Scientific and Industrial Research , Chandigarh , India.

Chemical modifications have been extensively exploited to circumvent shortcomings

in therapeutic applications of small interfering RNAs (siRNAs). However,

experimental designing and testing of these siRNAs or chemically modified siRNAs

(cm-siRNAs) involves enormous resources. Therefore, in-silico intervention in

designing cm-siRNAs would be of utmost importance. We developed SMEpred workbench

to predict the efficacy of normal siRNAs as well as cm-siRNAs using 3031

heterogeneous cm-siRNA sequences from siRNAmod database. These include 30

frequently used chemical modifications on different positions of either siRNA

strand. Support Vector Machine (SVM) was employed to develop predictive models

utilizing various sequence features namely mono-, di-nucleotide composition,

binary pattern and their hybrids. We achieved highest Pearson Correlation

Coefficient (PCC) of 0.80 during 10-fold cross validation and similar PCC value

in independent validation. We have provided the algorithm in the 'SMEpred'

pipeline to predict the normal siRNAs from the gene or mRNA sequence. For

multiple modifications, we have assembled 'MultiModGen' module to design multiple

modifications and further process them to evaluate their predicted efficacies.

SMEpred webserver will be useful to scientific community engaged in use of

RNAi-based technology as well as for therapeutic development. Web server is

available for public use at following URL address:

http://bioinfo.imtech.res.in/manojk/smepred .

DOI: 10.1080/15476286.2016.1229733

PMCID: PMC5100349 [Available on 2017-09-07]

PMID: 27603513

80. Gene. 2016 Oct 30;592(1):227-34. doi: 10.1016/j.gene.2016.07.059. Epub 2016 Jul

25.

Gene expression classification using epigenetic features and DNA sequence

composition in the human embryonic stem cell line H1.

Su WX(1), Li QZ(2), Zhang LQ(1), Fan GL(1), Wu CY(1), Yan ZH(1), Zuo YC(3).

Author information:

(1)Laboratory of Theoretical Biophysics, School of Physical Science and

Technology, Inner Mongolia University, Hohhot 010021, China. (2)Laboratory of

Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia

University, Hohhot 010021, China. Electronic address: qzli@imu.edu.cn. (3)The Key

Laboratory of Mammalian Reproductive Biology and Biotechnology of the Ministry of

Education, College of life sciences, Inner Mongolia University, Hohhot, 010021,

China.

Epigenetic factors are known to correlate with gene expression in the existing

studies. However, quantitative models that accurately classify the highly and

lowly expressed genes based on epigenetic factors are currently lacking. In this

study, a new machine learning method combines histone modifications, DNA

methylation, DNA accessibility, transcription factors, and trinucleotide

composition with support vector machines (SVM) is developed in the context of

human embryonic stem cell line (H1). The results indicate that the predictive

accuracy will be markedly improved when the epigenetic features are considered.

The predictive accuracy and Matthews correlation coefficient of the best model

are as high as 95.96% and 0.92 for 10-fold cross-validation test, and 95.58% and

0.92 for independent dataset test, respectively. Our model provides a good way to

judge a gene is either highly or lowly expressed gene by using genetic and

epigenetic data, when the expression data of the gene is lacking. And a

web-server GECES for our analysis method is established at

http://202.207.14.87:8032/fuwu/GECES/index.asp, so that other scientists can

easily get their desired results by our web-server, without going through the

mathematical details.

Copyright © 2016 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.gene.2016.07.059

PMID: 27468948 [Indexed for MEDLINE]

81. Bioinformatics. 2016 Oct 25. pii: btw652. [Epub ahead of print]

Isomorphic semantic mapping of variant call format (VCF2RDF).

Penha ED(1), Iriabho E(1), Dussaq A(1), de Oliveira DM(2), Almeida JS(3).

Author information:

(1)Department of Pathology, Informatics Division, University of Alabama at

Birmingham, Birmingham, AL 35233, USA. (2)Rede Nordeste de Biotecnologia,

Universidade Estadual do Ceará, Fortaleza CE 60740-000, Brazil. (3)Department of

Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794, USA.

The move of computational genomics workflows to Cloud Computing platforms is

associated with a new level of integration and interoperability that challenges

existing data representation formats. The Variant Calling Format (VCF) is in a

particularly sensitive position in that regard, with both clinical and

consumer-facing analysis tools relying on this self-contained description of

genomic variation in Next Generation Sequencing (NGS) results. In this report we

identify an isomorphic map between VCF and the reference Resource Description

Framework. RDF is advanced by the World Wide Web Consortium (W3C) to enable

representations of linked data that are both distributed and discoverable. The

resulting ability to decompose VCF reports of genomic variation without loss of

context addresses the need to modularize and govern NGS pipelines for Precision

Medicine. Specifically, it provides the flexibility (i.e. the indexing) needed to

support the wide variety of clinical scenarios and patient-facing governance

where only part of the VCF data is fitting.AVAILABILITY AND IMPLEMENTATION:

Software libraries with a claim to be both domain-facing and consumer-facing have

to pass the test of portability across the variety of devices that those

consumers in fact adopt. That is, ideally the implementation should itself take

place within the space defined by web technologies. Consequently, the isomorphic

mapping function was implemented in JavaScript, and was tested in a variety of

environments and devices, client and server side alike. These range from web

browsers in mobile phones to the most popular micro service platform, NodeJS. The

code is publicly available at https://github.com/ibl/VCFr, with a live deployment

at: http://ibl.github.io/VCFr/ CONTACT: jonas.almeida@stonybrookmedicine.edu.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw652

PMID: 27797761

82. Oncotarget. 2016 Oct 25;7(43):69783-69793. doi: 10.18632/oncotarget.11975.

iOri-Human: identify human origin of replication by incorporating dinucleotide

physicochemical properties into pseudo nucleotide composition.

Zhang CJ(1), Tang H(2), Li WC(1), Lin H(1,)(3), Chen W(1,)(4,)(3), Chou

KC(1,)(4,)(3).

Author information:

(1)Key Laboratory for Neuro-Information of Ministry of Education, School of Life

Science and Technology, Center for Informational Biology, University of

Electronic Science and Technology of China, Chengdu, 610054, China. (2)Department

of Pathophysiology, Southwest Medical University, Luzhou, 646000, China.

(3)Gordon Life Science Institute, Boston, MA, 02478, USA. (4)Department of

Physics, School of Sciences, and Center for Genomics and Computational Biology,

North China University of Science and Technology, Tangshan, 063000, China.

The initiation of replication is an extremely important process in DNA life

cycle. Given an uncharacterized DNA sequence, can we identify where its origin of

replication (ORI) is located? It is no doubt a fundamental problem in genome

analysis. Particularly, with the rapid development of genome sequencing

technology that results in a huge amount of sequence data, it is highly desired

to develop computational methods for rapidly and effectively identifying the ORIs

in these genomes. Unfortunately, by means of the existing computational methods,

such as sequence alignment or kmer strategies, it could hardly achieve decent

success rates. To address this problem, we developed a predictor called

"iOri-Human". Rigorous jackknife tests have shown that its overall accuracy and

stability in identifying human ORIs are over 75% and 50%, respectively. In the

predictor, it is through the pseudo nucleotide composition (an extension of

pseudo amino acid composition) that 96 physicochemical properties for the 16

possible constituent dinucleotides have been incorporated to reflect the global

sequence patterns in DNA as well as its local sequence patterns. Moreover, a

user-friendly web-server for iOri-Human has been established at

http://lin.uestc.edu.cn/server/iOri-Human.html, by which users can easily get

their desired results without the need to through the complicated mathematics

involved.

DOI: 10.18632/oncotarget.11975

PMID: 27626500

83. J Chem Inf Model. 2016 Oct 24;56(10):2115-2122. Epub 2016 Sep 22.

Sequence-Based Prediction of Protein-Carbohydrate Binding Sites Using Support

Vector Machines.

Taherzadeh G(1), Zhou Y(1), Liew AW(1), Yang Y(1).

Author information:

(1)School of Information and Communication Technology and ‡Institute for

Glycomics, Griffith University , Parklands Drive, Southport, Queensland 4215,

Australia.

Carbohydrate-binding proteins play significant roles in many diseases including

cancer. Here, we established a machine-learning-based method (called

sequence-based prediction of residue-level interaction sites of carbohydrates,

SPRINT-CBH) to predict carbohydrate-binding sites in proteins using support

vector machines (SVMs). We found that integrating evolution-derived sequence

profiles with additional information on sequence and predicted solvent accessible

surface area leads to a reasonably accurate, robust, and predictive method, with

area under receiver operating characteristic curve (AUC) of 0.78 and 0.77 and

Matthew's correlation coefficient of 0.34 and 0.29, respectively for 10-fold

cross validation and independent test without balancing binding and nonbinding

residues. The quality of the method is further demonstrated by having

statistically significantly more binding residues predicted for

carbohydrate-binding proteins than presumptive nonbinding proteins in the human

proteome, and by the bias of rare alleles toward predicted carbohydrate-binding

sites for nonsynonymous mutations from the 1000 genome project. SPRINT-CBH is

available as an online server at http://sparks-lab.org/server/SPRINT-CBH .

DOI: 10.1021/acs.jcim.6b00320

PMID: 27623166

84. Bioinformatics. 2016 Oct 22. pii: btw656. [Epub ahead of print]

SChloro: directing Viridiplantae proteins to six chloroplastic sub-compartments.

Savojardo C(1), Martelli PL(1), Fariselli P(2), Casadio R(1,)(3).

Author information:

(1)Biocomputing Group, BiGeA - CIG, Interdepartmental Center «Luigi Galvani» for

Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of

Bologna, Bologna, Italy. (2)Department of Comparative Biomedicine and Food

Science (BCA), University of Padova, Padova, Italy. (3)Interdepartmental Center

«Giorgio Prodi» for Cancer Research, University of Bologna, Bologna, Italy.

MOTIVATION: Chloroplasts are organelles found in plants and involved in several

important cell processes. Similarly to other compartments in the cell,

chloroplasts have an internal structure comprising several sub-compartments,

where different proteins are targeted to perform their functions. Given the

relation between protein function and localization, the availability of effective

computational tools to predict protein sub-organelle localizations is crucial for

large-scale functional studies.

RESULTS: In this paper we present SChloro, a novel machine-learning approach to

predict protein sub-chloroplastic localization, based on targeting signal

detection and membrane protein information. The proposed approach performs

multi-label predictions discriminating six chloroplastic sub-compartments that

include inner membrane, outer membrane, stroma, thylakoid lumen, plastoglobule

and thylakoid membrane. In comparative benchmarks, the proposed method

outperforms current state-of-the-art methods in both single- and

multi-compartment predictions, with an overall multi-label accuracy of 74%. The

results demonstrate the relevance of the approach that is eligible as a good

candidate for integration into more general large-scale annotation pipelines of

protein subcellular localization.

AVAILABILITY AND IMPLEMENTATION: The method is available as web server at

http://schloro.biocomp.unibo.it CONTACT: gigi@biocomp.unibo.it.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw656

PMID: 27797765

85. Bioinformatics. 2016 Oct 15;32(20):3133-3141. Epub 2016 Jun 26.

pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant

algorithm by incorporating sequence-coupled effects into general PseAAC.

Jia J(1), Zhang L(2), Liu Z(3), Xiao X(1), Chou KC(4).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China Gordon Life Science Institute, Boston, MA 02478, USA. (2)Computer

Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China. (3)School

of Computer Science and Engineering, Nanjing University of Science and

Technology, Nanjing 210094, China. (4)Gordon Life Science Institute, Boston, MA

02478, USA Center of Excellence in Genomic Medicine Research (CEGMR), King

Abdulaziz University, Jeddah 21589, Saudi Arabia Center of Bioinformatics, School

of Life Science and Technology, University of Electronic Science and Technology

of China, Chengdu 610054, China.

MOTIVATION: Sumoylation is a post-translational modification (PTM) process, in

which small ubiquitin-related modifier (SUMO) is attaching by covalent bonds to

substrate protein. It is critical to many different biological processes such as

replicating genome, expressing gene, localizing and stabilizing proteins;

unfortunately, it is also involved with many major disorders including

Alzheimer's and Parkinson's diseases. Therefore, for both basic research and drug

development, it is important to identify the sumoylation sites in proteins.

RESULTS: To address such a problem, we developed a predictor called pSumo-CD by

incorporating the sequence-coupled information into the general pseudo-amino acid

composition (PseAAC) and introducing the covariance discriminant (CD) algorithm,

in which a bias-adjustment term, which has the function to automatically adjust

the errors caused by the bias due to the imbalance of training data, had been

incorporated. Rigorous cross-validations indicated that the new predictor

remarkably outperformed the existing state-of-the-art prediction method for the

same purpose.

AVAILABILITY AND IMPLEMENTATION: For the convenience of most experimental

scientists, a user-friendly web-server for pSumo-CD has been established at

http://www.jci-bioinfo.cn/pSumo-CD, by which users can easily obtain their

desired results without the need to go through the complicated mathematical

equations involved.

CONTACT: jjia@gordonlifescience.org, xxiao@gordonlifescience.org or

kcchou@gordonlifescience.orgSupplementary information: Supplementary data are

available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw387

PMID: 27354696

86. Bioinformatics. 2016 Oct 15;32(20):3116-3123. Epub 2016 Jun 22.

iPTM-mLys: identifying multiple lysine PTM sites and their different types.

Qiu WR(1), Sun BQ(2), Xiao X(3), Xu ZC(2), Chou KC(4).

Author information:

(1)Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333403, China,

Department of Computer Science and Bond Life Science Center, University of

Missouri, Columbia MO, USA; Computational Biology, Gordon Life Science Institute,

Boston, MA 02478, USA. (2)Computer Department, Jingdezhen Ceramic Institute,

Jingdezhen 333403, China. (3)Computer Department, Jingdezhen Ceramic Institute,

Jingdezhen 333403, China, Computational Biology, Gordon Life Science Institute,

Boston, MA 02478, USA. (4)Computational Biology, Gordon Life Science Institute,

Boston, MA 02478, USA, Center of Bioinformatics, School of Life Science and

Technology, University of Electronic Science and Technology of China, Chengdu

610054, China Center of Excellence in Genomic Medicine Research (CEGMR), King

Abdulaziz University, Jeddah 21589, Saudi Arabia.

MOTIVATION: Post-translational modification, abbreviated as PTM, refers to the

change of the amino acid side chains of a protein after its biosynthesis. Owing

to its significance for in-depth understanding various biological processes and

developing effective drugs, prediction of PTM sites in proteins have currently

become a hot topic in bioinformatics. Although many computational methods were

established to identify various single-label PTM types and their occurrence sites

in proteins, no method has ever been developed for multi-label PTM types. As one

of the most frequently observed PTMs, the K-PTM, namely, the modification

occurring at lysine (K), can be usually accommodated with many different types,

such as 'acetylation', 'crotonylation', 'methylation' and 'succinylation'. Now we

are facing an interesting challenge: given an uncharacterized protein sequence

containing many K residues, which ones can accommodate two or more types of PTM,

which ones only one, and which ones none?

RESULTS: To address this problem, a multi-label predictor called IPTM-MLYS: has

been developed. It represents the first multi-label PTM predictor ever

established. The novel predictor is featured by incorporating the

sequence-coupled effects into the general PseAAC, and by fusing an array of basic

random forest classifiers into an ensemble system. Rigorous cross-validations via

a set of multi-label metrics indicate that the first multi-label PTM predictor is

very promising and encouraging.

AVAILABILITY AND IMPLEMENTATION: For the convenience of most experimental

scientists, a user-friendly web-server for iPTM-mLys has been established at

http://www.jci-bioinfo.cn/iPTM-mLys, by which users can easily obtain their

desired results without the need to go through the complicated mathematical

equations involved.

CONTACT: wqiu@gordonlifescience.org, xxiao@gordonlifescience.org,

kcchou@gordonlifescience.orgSupplementary information: Supplementary data are

available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw380

PMID: 27334473

87. Mar Pollut Bull. 2016 Oct 15;111(1-2):330-8. doi:

10.1016/j.marpolbul.2016.06.090. Epub 2016 Jul 4.

Fate, behaviour and weathering of priority HNS in the marine environment: An

online tool.

Cunha I(1), Oliveira H(2), Neuparth T(1), Torres T(1), Santos MM(3).

Author information:

(1)CIMAR/CIIMAR, Interdisciplinary Centre of Marine and Environmental Research,

University of Porto, Rua dos Bragas 289, 4050-123 Porto, Portugal.

(2)CIMAR/CIIMAR, Interdisciplinary Centre of Marine and Environmental Research,

University of Porto, Rua dos Bragas 289, 4050-123 Porto, Portugal. Electronic

address: helaoliveira@gmail.com. (3)CIMAR/CIIMAR, Interdisciplinary Centre of

Marine and Environmental Research, University of Porto, Rua dos Bragas 289,

4050-123 Porto, Portugal; FCUP, Department of Biology, Faculty of Sciences,

University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal. Electronic

address: santos@ciimar.up.pt.

Literature data and data obtained with modelling tools were compiled to derive

the physicochemical behaviour of 24 priority Hazardous and Noxious Substances

(HNS), as a proxy to improve environmental, public health and political issues in

relation to HNS spills. Parameters that rule the HNS behaviour in water and those

that determine their distribution and persistence in the environment, such as

fugacity, physicochemical degradation, biodegradation,

bioaccumulation/biotransformation and aquatic toxicity, were selected. Data

systematized and produced in the frame of the Arcopol Platform project was made

available through a public database (http://www.ciimar.up.pt/hns/substances.php).

This tool is expected to assist stakeholders involved in HNS spills preparedness

and response, policy makers and legislators, as well as to contribute to a

current picture of the scientific knowledge on the fate, behaviour, weathering

and toxicity of priority HNS, being essential to support future improvements in

maritime safety and coastal pollution response before, during and after spill

incidents.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.marpolbul.2016.06.090

PMID: 27389461

88. Bioinformatics. 2016 Oct 14. pii: btw639. [Epub ahead of print]

LDAP: a web server for lncRNA-disease association prediction.

Lan W(1), Li M(1), Zhao K(1), Liu J(1), Wu FX(2), Pan Y(3), Wang J(1).

Author information:

(1)School of Information Science and Engineering, Central South University,

Changsha 410083, China. (2)Division of Biomedical Engineering, University of

Saskatchewan, Saskatoon SKS7N5A9, Canada. (3)Department of Computer Science,

Georgia State University, Atlanta, GA 30302, USA.

MOTIVATION: Increasing evidences have demonstrated that long noncoding RNAs

(lncRNAs) play important roles in many human diseases. Therefore, predicting

novel lncRNA-disease associations would contribute to dissect the complex

mechanisms of disease pathogenesis. Some computational methods have been

developed to infer lncRNA-disease associations. However, most of these methods

infer lncRNA-disease associations only based on single data resource.

RESULTS: In this paper, we propose a new computational method to predict

lncRNA-disease associations by integrating multiple biological data resources.

Then, we implement this method as a web server for lncRNA-disease association

prediction (LDAP). The input of the LDAP server is the lncRNA sequence. The LDAP

predicts potential lncRNA-disease associations by using a bagging SVM classifier

based on lncRNA similarity and disease similarity.

AVAILABILITY AND IMPLEMENTATION: The web server is available at

http://bioinformatics.csu.edu.cn/ldap CONTACT:

jxwang@mail.csu.edu.cnSupplementary information: Supplementary data are available

at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw639

PMID: 27742699

89. Bioinformatics. 2016 Oct 14. pii: btw644. [Epub ahead of print]

iATC-mISF: a multi-label classifier for predicting the classes of anatomical

therapeutic chemicals.

Cheng X(1,)(2), Zhao SG(1), Xiao X(2,)(3), Chou KC(3,)(4,)(5).

Author information:

(1)College of Information Science and Technology, Donghua University, Shanghai,

China. (2)Computer Department, Jingdezhen Ceramic Institute, Jingdezhen, China.

(3)Gordon Life Science Institute, Boston, MA 02478, USA. (4)Key Laboratory for

Neuro-Information of Ministry of Education, School of Life Science and

Technology, Center for Informational Biology, University of Electronic Science

and Technology of China, Chengdu 610054, China. (5)Center of Excellence in

Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi

Arabia.

MOTIVATION: Given a compound, can we predict which anatomical therapeutic

chemical (ATC) class/classes it belongs to? It is a challenging problem since the

information thus obtained can be used to deduce its possible active ingredients,

as well as its therapeutic, pharmacological and chemical properties. And hence

the pace of drug development could be substantially expedited. But this problem

is by no means an easy one. Particularly, some drugs or compounds may belong to

two or more ATC classes.

RESULTS: To address it, a multi-label classifier, called IATC-MISF: , was

developed by incorporating the information of chemical-chemical interaction, the

information of the structural similarity, and the information of the

fingerprintal similarity. Rigorous cross-validations showed that the proposed

predictor achieved remarkably higher prediction quality than its cohorts for the

same purpose, particularly in the absolute true rate, the most important and

harsh metrics for the multi-label systems.

AVAILABILITY AND IMPLEMENTATION: The web-server for iATC-mISF is accessible at

http://www.jci-bioinfo.cn/iATC-mISF Furthermore, to maximize the convenience for

most experimental scientists, a step-by-step guide was provided, by which users

can easily get their desired results without needing to go through the

complicated mathematical equations. Their inclusion in this article is just for

the integrity of the new method and stimulating more powerful methods to deal

with various multi-label systems in biology.

CONTACT: xxiao@gordonlifescience.orgSupplementary information: Supplementary data

are available at Bioinformatics online.

© The Authors 2016. Published by Oxford University Press. All rights reserved.

For Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw644

PMID: 27742696

90. Sci Rep. 2016 Oct 13;6:34651. doi: 10.1038/srep34651.

3Disease Browser: A Web server for integrating 3D genome and disease-associated

chromosome rearrangement data.

Li R(1), Liu Y(2), Li T(1), Li C(1,)(3).

Author information:

(1)Peking-Tsinghua Center for Life Sciences, Academy for Advanced

Interdisciplinary Studies; School of Life Sciences, Peking University, Beijing,

China. (2)School of Life Sciences, Tsinghua University, Beijing, China. (3)Center

for Statistical Science; Center for Bioinformatics, Peking University, Beijing,

China.

Chromosomal rearrangement (CR) events have been implicated in many tumor and

non-tumor human diseases. CR events lead to their associated diseases by

disrupting gene and protein structures. Also, they can lead to diseases through

changes in chromosomal 3D structure and gene expression. In this study, we search

for CR-associated diseases potentially caused by chromosomal 3D structure

alteration by integrating Hi-C and ChIP-seq data. Our algorithm rediscovers

experimentally verified disease-associated CRs (polydactyly diseases) that alter

gene expression by disrupting chromosome 3D structure. Interestingly, we find

that intellectual disability may be a candidate disease caused by 3D chromosome

structure alteration. We also develop a Web server (3Disease Browser,

http://3dgb.cbi.pku.edu.cn/disease/) for integrating and visualizing

disease-associated CR events and chromosomal 3D structure.

DOI: 10.1038/srep34651

PMCID: PMC5062081

PMID: 27734896

91. IEEE/ACM Trans Comput Biol Bioinform. 2016 Oct 11. [Epub ahead of print]

Predicting Protein-DNA Binding Residues by Weightedly Combining Sequence-based

Features and Boosting Multiple SVMs.

Hu J, Li Y, Zhang M, Yang X, Shen HB, Yu DJ.

Protein-DNA interactions are ubiquitous in a wide variety of biological

processes. Correctly locating DNA-binding residues solely from protein sequences

is an important but challenging task for protein function annotations and drug

discovery, especially in the post-genomic era where large volumes of protein

sequences have quickly accumulated. In this study, we report a new predictor,

named TargetDNA, for targeting protein-DNA binding residues from primary

sequences. TargetDNA uses a protein's evolutionary information and its predicted

solvent accessibility as two base features and employs a centered linear kernel

alignment algorithm to learn the weights for weightedly combining the two

features. Based on the weightedly combined feature, multiple initial predictors

with SVM as classifiers are trained by applying a random under-sampling technique

to the original dataset, the purpose of which is to cope with the severe

imbalance phenomenon that exists between the number of DNA-binding and

non-binding residues. The final ensembled predictor is obtained by boosting the

multiple initially trained predictors. Experimental simulation results

demonstrate that the proposed TargetDNA achieves a high prediction performance

and outperforms many existing sequence-based protein-DNA binding residue

predictors. The TargetDNA web server and datasets are freely available at

http://csbio.njust.edu.cn/bioinf/TargetDNA/ for academic use.

DOI: 10.1109/TCBB.2016.2616469

PMID: 27740495

92. Sci Rep. 2016 Oct 11;6:35123. doi: 10.1038/srep35123.

PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide

compositions.

Chen W(1), Feng P(2), Ding H(3), Lin H(3).

Author information:

(1)Department of Physics, School of Sciences, and Center for Genomics and

Computational Biology, North China University of Science and Technology,

Tangshan, 063000, China. (2)School of Public Health, North China University of

Science and Technology, Tangshan, 063000, China. (3)Key Laboratory for

Neuro-Information of Ministry of Education, School of Life Science and

Technology, Center for Informational Biology, University of Electronic Science

and Technology of China, Chengdu 610054, China.

The adenosine to inosine (A-to-I) editing is the most prevalent kind of RNA

editing and involves in many biological processes. Accurate identification of

A-to-I editing site is invaluable for better understanding its biological

functions. Due to the limitations of experimental methods, in the present study,

a support vector machine based-model, called PAI, is proposed to identify A-to-I

editing site in D. melanogaster. In this model, RNA sequences are encoded by

"pseudo dinucleotide composition" into which six RNA physiochemical properties

were incorporated. PAI achieves promising performances in jackknife test and

independent dataset test, indicating that it holds very high potential to become

a useful tool for identifying A-to-I editing site. For the convenience of

experimental scientists, a web-server was constructed for PAI and it is freely

accessible at http://lin.uestc.edu.cn/server/PAI.

DOI: 10.1038/srep35123

PMCID: PMC5057124

PMID: 27725762

93. Sci Rep. 2016 Oct 11;6:34985. doi: 10.1038/srep34985.

circRNADb: A comprehensive database for human circular RNAs with protein-coding

annotations.

Chen X(1), Han P(2), Zhou T(3), Guo X(3), Song X(1), Li Y(3,)(4).

Author information:

(1)Department of Biomedical Engineering, Nanjing University of Aeronautics and

Astronautics, Nanjing 211106, China. (2)Department of Gynecology and Obstetrics,

The First Affiliated Hospital with Nanjing Medical University, Nanjing 210029,

China. (3)State Key Laboratory of Reproductive Medicine, Department of Histology

and Embryology, Nanjing Medical University, Nanjing 210029, China. (4)Center of

Pathology and Clinical Laboratory, Sir Run Run Hospital Affiliated with Nanjing

Medical University, Nanjing 211166, China.

It has been known that circular RNAs are widely expressed in human tissues and

cells, and play important regulatory roles in physiological or pathological

processes. However, there is lack of comprehensively annotated human circular

RNAs database. In this study we established a circRNA database, named as

circRNADb, containing 32,914 human exonic circRNAs carefully selected from

diversified sources. The detailed information of the circRNA, including genomic

information, exon splicing, genome sequence, internal ribosome entry site (IRES),

open reading frame (ORF) and references were provided in circRNADb. In addition,

circRNAs were found to be able to encode proteins, which have not been reported

in any species. 16328 circRNAs were annotated to have ORF longer than 100 amino

acids, of which 7170 have IRES elements. 46 circRNAs from 37 genes were found to

have their corresponding proteins expressed according mass spectrometry. The

database provides the function of data search, browse, download, submit and

feedback for the user to study particular circular RNA of interest and update the

database continually. circRNADb will be built to be a biological information

platform for circRNA molecules and related biological functions in the future.

The database can be freely available through the web server at

http://reprod.njmu.edu.cn/circrnadb.

DOI: 10.1038/srep34985

PMCID: PMC5057092

PMID: 27725737

94. PLoS One. 2016 Oct 10;11(10):e0162707. doi: 10.1371/journal.pone.0162707.

eCollection 2016.

RNAMethPre: A Web Server for the Prediction and Query of mRNA m6A Sites.

Xiang S(1), Liu K(1,)(2), Yan Z(1), Zhang Y(3), Sun Z(1).

Author information:

(1)MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua

University, Beijing, 100084, P. R. China. (2)Department of Statistics, University

of California, Berkeley, California, 94720, United States of America. (3)Key Lab

in Healthy Science and Technology, Division of Life Science, Graduate School at

Shenzhen, Tsinghua University, Beijing, 518055, P. R. China.

N6-Methyladenosine (m6A) is the most common mRNA modification; it occurs in a

wide range of taxon and is associated with many key biological processes.

High-throughput experiments have identified m6A-peaks and sites across the

transcriptome, but studies of m6A sites at the transcriptome-wide scale are

limited to a few species and tissue types. Therefore, the computational

prediction of mRNA m6A sites has become an important strategy. In this study, we

integrated multiple features of mRNA (flanking sequences, local secondary

structure information, and relative position information) and trained a SVM

classifier to predict m6A sites in mammalian mRNA sequences. Our method achieves

ideal performance in both cross-validation tests and rigorous independent dataset

tests. The server also provides a comprehensive database of predicted

transcriptome-wide m6A sites and curated m6A-seq peaks from the literature for

both human and mouse, and these can be queried and visualized in a genome

browser. The RNAMethPre web server provides a user-friendly tool for the

prediction and query of mRNA m6A sites, which is freely accessible for public use

at http://bioinfo.tsinghua.edu.cn/RNAMethPre/index.html.

DOI: 10.1371/journal.pone.0162707

PMCID: PMC5056760

PMID: 27723837

Conflict of interest statement: The authors have declared that no competing

interests exist.

95. BMC Bioinformatics. 2016 Oct 7;17(1):411.

RStrucFam: a web server to associate structure and cognate RNA for RNA-binding

proteins from sequence information.

Ghosh P(1), Mathew OK(1,)(2), Sowdhamini R(3).

Author information:

(1)National Centre for Biological Sciences, Tata Institute of Fundamental

Research, Bellary Road, Bangalore, Karnataka, 560 065, India. (2)SASTRA

University, Tirumalaisamudram, Thanjavur, 613401, Tamil Nadu, India. (3)National

Centre for Biological Sciences, Tata Institute of Fundamental Research, Bellary

Road, Bangalore, Karnataka, 560 065, India. mini@ncbs.res.in.

BACKGROUND: RNA-binding proteins (RBPs) interact with their cognate RNA(s) to

form large biomolecular assemblies. They are versatile in their functionality and

are involved in a myriad of processes inside the cell. RBPs with similar

structural features and common biological functions are grouped together into

families and superfamilies. It will be useful to obtain an early understanding

and association of RNA-binding property of sequences of gene products. Here, we

report a web server, RStrucFam, to predict the structure, type of cognate RNA(s)

and function(s) of proteins, where possible, from mere sequence information.

RESULTS: The web server employs Hidden Markov Model scan (hmmscan) to enable

association to a back-end database of structural and sequence families. The

database (HMMRBP) comprises of 437 HMMs of RBP families of known structure that

have been generated using structure-based sequence alignments and 746

sequence-centric RBP family HMMs. The input protein sequence is associated with

structural or sequence domain families, if structure or sequence signatures

exist. In case of association of the protein with a family of known structures,

output features like, multiple structure-based sequence alignment (MSSA) of the

query with all others members of that family is provided. Further, cognate RNA

partner(s) for that protein, Gene Ontology (GO) annotations, if any and a

homology model of the protein can be obtained. The users can also browse through

the database for details pertaining to each family, protein or RNA and their

related information based on keyword search or RNA motif search.

CONCLUSIONS: RStrucFam is a web server that exploits structurally conserved

features of RBPs, derived from known family members and imprinted in mathematical

profiles, to predict putative RBPs from sequence information. Proteins that fail

to associate with such structure-centric families are further queried against the

sequence-centric RBP family HMMs in the HMMRBP database. Further, all other

essential information pertaining to an RBP, like overall function annotations,

are provided. The web server can be accessed at the following link:

http://caps.ncbs.res.in/rstrucfam .

DOI: 10.1186/s12859-016-1289-x

PMCID: PMC5054549

PMID: 27717309

96. IEEE/ACM Trans Comput Biol Bioinform. 2016 Oct 4. doi: 10.1109/TCBB.2016.2615010.

[Epub ahead of print]

HEMEsPred: Structure-based Ligand-specific Heme Binding Residues Prediction by

Using Fast-adaptive Ensemble Learning Scheme.

Zhang J, Chai H, Gao B, Yang G, Ma Z.

Heme is an essential biomolecule that widely exists in numerous extant organisms.

Accurately identifying heme binding residues (HEMEs) is of great importance in

disease progression and drug development. In this study, a novel predictor named

HEMEsPred was proposed for predicting HEMEs. First, several sequence- and

structure-based features, including amino acid composition, motifs, surface

preferences and secondary structure, were collected to construct feature

matrices. Second, a novel fast-adaptive ensemble learning scheme was designed to

overcome the serious class-imbalance problem as well as to enhance the prediction

performance. Third, we further developed ligand-specific models considering that

different heme ligands varied significantly in their roles, sizes and

distributions. Statistical test proved the effectiveness of ligand-specific

models. Experimental results on benchmark datasets demonstrated good robustness

of our proposed method. Furthermore, our method also showed good generalization

capability and outperformed many state-of-art predictors on two independent

testing datasets. HEMEsPred web server was available at

http://www.inforstation.com/HEMEsPred/ for free academic use.

DOI: 10.1109/TCBB.2016.2615010

PMID: 28029626

97. IEEE/ACM Trans Comput Biol Bioinform. 2016 Oct 4. [Epub ahead of print]

HEMEsPred: Structure-based Ligand-specific Heme Binding Residues Prediction by

Using Fast-adaptive Ensemble Learning Scheme.

Zhang J, Chai H, Gao B, Yang G, Ma Z.

Heme is an essential biomolecule that widely exists in numerous extant organisms.

Accurately identifying heme binding residues (HEMEs) is of great importance in

disease progression and drug development. In this study, a novel predictor named

HEMEsPred was proposed for predicting HEMEs. First, several sequence- and

structure-based features, including amino acid composition, motifs, surface

preferences and secondary structure, were collected to construct feature

matrices. Second, a novel fast-adaptive ensemble learning scheme was designed to

overcome the serious class-imbalance problem as well as to enhance the prediction

performance. Third, we further developed ligand-specific models considering that

different heme ligands varied significantly in their roles, sizes and

distributions. Statistical test proved the effectiveness of ligand-specific

models. Experimental results on benchmark datasets demonstrated good robustness

of our proposed method. Furthermore, our method also showed good generalization

capability and outperformed many state-of-art predictors on two independent

testing datasets. HEMEsPred web server was available at

http://59.73.198.144:8080/HEMEsPred/ for free academic use.

DOI: 10.1109/TCBB.2016.2615010

PMID: 27775533

98. PLoS One. 2016 Oct 3;11(10):e0163497. doi: 10.1371/journal.pone.0163497.

eCollection 2016.

Graphlet Based Metrics for the Comparison of Gene Regulatory Networks.

Martin AJ(1,)(2), Dominguez C(1,)(3), Contreras-Riquelme S(1), Holmes DS(3),

Perez-Acle T(1,)(2).

Author information:

(1)Computational Biology Lab, Fundación Ciencia & Vida, Santiago, Chile.

(2)Centro Interdisciplinario de Neurociencia de Valparaíso, Universidad de

Valparaíso, Chile. (3)Center for Bioinformatics and Genome Biology, Fundación

Ciencia & Vida and Facultad de Ciencias Biologicas, Universidad Andres Bello,

Santiago, Chile.

Understanding the control of gene expression remains one of the main challenges

in the post-genomic era. Accordingly, a plethora of methods exists to identify

variations in gene expression levels. These variations underlay almost all

relevant biological phenomena, including disease and adaptation to environmental

conditions. However, computational tools to identify how regulation changes are

scarce. Regulation of gene expression is usually depicted in the form of a gene

regulatory network (GRN). Structural changes in a GRN over time and conditions

represent variations in the regulation of gene expression. Like other biological

networks, GRNs are composed of basic building blocks called graphlets. As a

consequence, two new metrics based on graphlets are proposed in this work:

REConstruction Rate (REC) and REC Graphlet Degree (RGD). REC determines the rate

of graphlet similarity between different states of a network and RGD identifies

the subset of nodes with the highest topological variation. In other words, RGD

discerns how th GRN was rewired. REC and RGD were used to compare the local

structure of nodes in condition-specific GRNs obtained from gene expression data

of Escherichia coli, forming biofilms and cultured in suspension. According to

our results, most of the network local structure remains unaltered in the two

compared conditions. Nevertheless, changes reported by RGD necessarily imply that

a different cohort of regulators (i.e. transcription factors (TFs)) appear on the

scene, shedding light on how the regulation of gene expression occurs when E.

coli transits from suspension to biofilm. Consequently, we propose that both

metrics REC and RGD should be adopted as a quantitative approach to conduct

differential analyses of GRNs. A tool that implements both metrics is available

as an on-line web server (http://dlab.cl/loto).

DOI: 10.1371/journal.pone.0163497

PMCID: PMC5047442

PMID: 27695050

Conflict of interest statement: The authors have declared that no competing

interests exist.

99. Bioinformatics. 2016 Oct 1;32(19):3041-3. doi: 10.1093/bioinformatics/btw332.

Epub 2016 Jun 13.

Web-based network analysis and visualization using CellMaps.

Salavert F(1), García-Alonso L(2), Sánchez R(2), Alonso R(3), Bleda M(4), Medina

I(5), Dopazo J(6).

Author information:

(1)Computational Genomics Department, Centro de Investigación Príncipe Felipe

(CIPF), Valencia 46012, Spain Bioinformatics of Rare Diseases (BIER), CIBER de

Enfermedades Raras (CIBERER), Valencia, Spain. (2)Computational Genomics

Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia 46012,

Spain. (3)Computational Genomics Department, Centro de Investigación Príncipe

Felipe (CIPF), Valencia 46012, Spain Bull-CIPF, Computational Genomics Chair,

Valencia 46012, Spain. (4)Department of Medicine, University of Cambridge, School

of Clinical Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK.

(5)HPC Service, UIS, University of Cambridge, Cambridge, UK. (6)Computational

Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia

46012, Spain Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras

(CIBERER), Valencia, Spain Functional Genomics Node, (INB, PRB2, ISCIII) at CIPF,

Valencia 46012, Spain.

: CellMaps is an HTML5 open-source web tool that allows displaying, editing,

exploring and analyzing biological networks as well as integrating metadata into

them. Computations and analyses are remotely executed in high-end servers, and

all the functionalities are available through RESTful web services. CellMaps can

easily be integrated in any web page by using an available JavaScript

API.AVAILABILITY AND IMPLEMENTATION: The application is available at:

http://cellmaps.babelomics.org/ and the code can be found in:

https://github.com/opencb/cell-maps The client is implemented in JavaScript and

the server in C and Java.

CONTACT: jdopazo@cipf.es

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw332

PMCID: PMC5039919

PMID: 27296979

100. Comput Methods Programs Biomed. 2016 Oct;135:101-4. doi:

10.1016/j.cmpb.2016.07.023. Epub 2016 Jul 21.

Digital questionnaire platform in the Danish Blood Donor Study.

Burgdorf KS(1), Felsted N(2), Mikkelsen S(3), Nielsen MH(4), Thørner LW(4),

Pedersen OB(5), Sørensen E(4), Nielsen KR(6), Bruun MT(7), Werge T(2), Erikstrup

C(3), Hansen T(2), Ullum H(4).

Author information:

(1)Department of Clinical Immunology, Copenhagen University Hospital,

Rigshospitalet, Copenhagen, Denmark. Electronic address:

kristoffer.soelvsten.burgdorf@regionh.dk. (2)Institute of Biological Psychiatry,

Mental Health Centre Sct. Hans, Copenhagen University Hospital, Roskilde,

Denmark. (3)Department of Clinical Immunology, Aarhus University Hospital,

Aarhus, Denmark. (4)Department of Clinical Immunology, Copenhagen University

Hospital, Rigshospitalet, Copenhagen, Denmark. (5)Department of Clinical

Immunology, Naestved Hospital, Naestved, Denmark. (6)Department of Clinical

Immunology, Aalborg University Hospital, Aalborg, Denmark. (7)Department of

Clinical Immunology, Odense University Hospital, Odense, Denmark.

OBJECTIVES: The Danish Blood Donor Study (DBDS) is a prospective,

population-based study and biobank. Since 2010, 100,000 Danish blood donors have

been included in the study. Prior to July 2015 all participating donors had to

complete a paper-based questionnaire. Here we describe the establishment of a

digital tablet-based questionnaire platform implemented in blood bank sites

across Denmark.

METHODS: The digital questionnaire was developed using the open source survey

software tool LimeSurvey. The participants accesses the questionnaire online with

a standard SSL encrypted HTTP connection using their personal civil registration

numbers. The questionnaire is placed at a front-end web server and a collection

server retrieves the completed questionnaires. Data from blood samples, register

data, genetic data and verification of signed informed consent are then

transferred to and merged with the questionnaire data in the DBDS database.

RESULTS: The digital platform enables personalized questionnaires, presenting

only questions relevant to the specific donor by hiding unneeded follow-up

questions on screening question results. New versions of questionnaires are

immediately available at all blood collection facilities when new projects are

initiated.

CONCLUSION: The digital platform is a faster, cost-effective and more flexible

solution to collect valid data from participating donors compared to paper-based

questionnaires. The overall system can be used around the world by the use of

Internet connection, but the level of security depends on the sensitivity of the

data to be collected.

Copyright © 2016 The Authors. Published by Elsevier Ireland Ltd.. All rights

reserved.

DOI: 10.1016/j.cmpb.2016.07.023

PMID: 27586483 [Indexed for MEDLINE]

101. IEEE Trans Nanobioscience. 2016 Oct;15(7):674-682. doi: 10.1109/TNB.2016.2599115.

Epub 2016 Aug 10.

TargetM6A: Identifying N(6)-Methyladenosine Sites From RNA Sequences via

Position-Specific Nucleotide Propensities and a Support Vector Machine.

Li GQ, Liu Z, Shen HB, Yu DJ.

As one of the most ubiquitous post-transcriptional modifications of RNA,

N(6)-methyladenosine ( [Formula: see text]) plays an essential role in many vital

biological processes. The identification of [Formula: see text] sites in RNAs is

significantly important for both basic biomedical research and practical drug

development. In this study, we designed a computational-based method, called

TargetM6A, to rapidly and accurately target [Formula: see text] sites solely from

the primary RNA sequences. Two new features, i.e., position-specific

nucleotide/dinucleotide propensities (PSNP/PSDP), are introduced and combined

with the traditional nucleotide composition (NC) feature to formulate RNA

sequences. The extracted features are further optimized to obtain a much more

compact and discriminative feature subset by applying an incremental feature

selection (IFS) procedure. Based on the optimized feature subset, we trained

TargetM6A on the training dataset with a support vector machine (SVM) as the

prediction engine. We compared the proposed TargetM6A method with existing

methods for predicting [Formula: see text] sites by performing stringent

jackknife tests and independent validation tests on benchmark datasets. The

experimental results show that the proposed TargetM6A method outperformed the

existing methods for predicting [Formula: see text] sites and remarkably improved

the prediction performances, with MCC = 0.526 and AUC = 0.818. We also provided a

user-friendly web server for TargetM6A, which is publicly accessible for academic

use at http://csbio.njust.edu.cn/bioinf/TargetM6A.

DOI: 10.1109/TNB.2016.2599115

PMID: 27552763

102. Protein Sci. 2016 Oct;25(10):1825-33. doi: 10.1002/pro.2991. Epub 2016 Aug 9.

Improving protein-protein interactions prediction accuracy using protein

evolutionary information and relevance vector machine model.

An JY(1), Meng FR(1), You ZH(2), Chen X(3), Yan GY(4), Hu JP(1).

Author information:

(1)School of Computer Science Technology, China University of Mining and

Technology, Xuzhou, Jiangsu, 21116, China. (2)School of Computer Science

Technology, China University of Mining and Technology, Xuzhou, Jiangsu, 21116,

China. zhuhongyou@hotmail.com. (3)School of Information and Electrical

Engineering, China University of Mining and Technology, Xuzhou, Jiangsu, 21116,

China. (4)Academy of Mathematics and Systems Science, Chinese Academy of

Sciences, Beijing, 100190, China.

Predicting protein-protein interactions (PPIs) is a challenging task and

essential to construct the protein interaction networks, which is important for

facilitating our understanding of the mechanisms of biological systems. Although

a number of high-throughput technologies have been proposed to predict PPIs,

there are unavoidable shortcomings, including high cost, time intensity, and

inherently high false positive rates. For these reasons, many computational

methods have been proposed for predicting PPIs. However, the problem is still far

from being solved. In this article, we propose a novel computational method

called RVM-BiGP that combines the relevance vector machine (RVM) model and

Bi-gram Probabilities (BiGP) for PPIs detection from protein sequences. The major

improvement includes (1) Protein sequences are represented using the Bi-gram

probabilities (BiGP) feature representation on a Position Specific Scoring Matrix

(PSSM), in which the protein evolutionary information is contained; (2) For

reducing the influence of noise, the Principal Component Analysis (PCA) method is

used to reduce the dimension of BiGP vector; (3) The powerful and robust

Relevance Vector Machine (RVM) algorithm is used for classification. Five-fold

cross-validation experiments executed on yeast and Helicobacter pylori datasets,

which achieved very high accuracies of 94.57 and 90.57%, respectively.

Experimental results are significantly better than previous methods. To further

evaluate the proposed method, we compare it with the state-of-the-art support

vector machine (SVM) classifier on the yeast dataset. The experimental results

demonstrate that our RVM-BiGP method is significantly better than the SVM-based

method. In addition, we achieved 97.15% accuracy on imbalance yeast dataset,

which is higher than that of balance yeast dataset. The promising experimental

results show the efficiency and robust of the proposed method, which can be an

automatic decision support tool for future proteomics research. For facilitating

extensive studies for future proteomics research, we developed a freely available

web server called RVM-BiGP-PPIs in Hypertext Preprocessor (PHP) for predicting

PPIs. The web server including source code and the datasets are available at

http://219.219.62.123:8888/BiGP/.

© 2016 The Authors Protein Science published by Wiley Periodicals, Inc. on behalf

of The Protein Society.

DOI: 10.1002/pro.2991

PMCID: PMC5029537

PMID: 27452983

103. Biochem Biophys Res Commun. 2016 Sep 30;478(4):1739-45. doi:

10.1016/j.bbrc.2016.09.015. Epub 2016 Sep 4.

Palindrome analyser - A new web-based server for predicting and evaluating

inverted repeats in nucleotide sequences.

Brázda V(1), Kolomazník J(2), Lýsek J(2), Hároníková L(3), Coufal J(3), Št'astný

J(2).

Author information:

(1)Institute of Biophysics, Academy of Sciences of the Czech Republic,

Královopolská 135, 612 65, Brno, Czech Republic. Electronic address:

vaclav@ibp.cz. (2)Mendel University in Brno, Zemědělská 1, 613 00, Brno, Czech

Republic. (3)Institute of Biophysics, Academy of Sciences of the Czech Republic,

Královopolská 135, 612 65, Brno, Czech Republic.

DNA cruciform structures play an important role in the regulation of natural

processes including gene replication and expression, as well as nucleosome

structure and recombination. They have also been implicated in the evolution and

development of diseases such as cancer and neurodegenerative disorders. Cruciform

structures are formed by inverted repeats, and their stability is enhanced by DNA

supercoiling and protein binding. They have received broad attention because of

their important roles in biology. Computational approaches to study inverted

repeats have allowed detailed analysis of genomes. However, currently there are

no easily accessible and user-friendly tools that can analyse inverted repeats,

especially among long nucleotide sequences. We have developed a web-based server,

Palindrome analyser, which is a user-friendly application for analysing inverted

repeats in various DNA (or RNA) sequences including genome sequences and

oligonucleotides. It allows users to search and retrieve desired gene/nucleotide

sequence entries from the NCBI databases, and provides data on length, sequence,

locations and energy required for cruciform formation. Palindrome analyser also

features an interactive graphical data representation of the distribution of the

inverted repeats, with options for sorting according to the length of inverted

repeat, length of loop, and number of mismatches. Palindrome analyser can be

accessed at http://bioinformatics.ibp.cz.

Copyright © 2016 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.bbrc.2016.09.015

PMID: 27603574

104. Nucleic Acids Res. 2016 Sep 30;44(17):8433-41. doi: 10.1093/nar/gkw558. Epub 2016

Jun 21.

RNAlien - Unsupervised RNA family model construction.

Eggenhofer F(1), Hofacker IL(2), Höner Zu Siederdissen C(3).

Author information:

(1)Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse

17, A-1090 Vienna, Austria Bioinformatics Group, Department of Computer Science

University of Freiburg, Georges-Köhler-Allee, 79110 Freiburg, Germany

egg@informatik.uni-freiburg.de. (2)Institute for Theoretical Chemistry,

University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria Research Group

Bioinformatics and Computational Biology, Faculty of Computer Science, University

of Vienna, A-1090 Vienna, Austria. (3)Institute for Theoretical Chemistry,

University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria Bioinformatics

Group, Department of Computer Science, University of Leipzig, D-04107 Leipzig,

Germany Interdisciplinary Center for Bioinformatics, University of Leipzig,

Härtelstraße 16-18, D-04107 Leipzig, Germany.

Determining the function of a non-coding RNA requires costly and time-consuming

wet-lab experiments. For this reason, computational methods which ascertain the

homology of a sequence and thereby deduce functionality and family membership are

often exploited. In this fashion, newly sequenced genomes can be annotated in a

completely computational way. Covariance models are commonly used to assign novel

RNA sequences to a known RNA family. However, to construct such models several

examples of the family have to be already known. Moreover, model building is the

work of experts who manually edit the necessary RNA alignment and consensus

structure. Our method, RNAlien, starting from a single input sequence collects

potential family member sequences by multiple iterations of homology search. RNA

family models are fully automatically constructed for the found sequences. We

have tested our method on a subset of the Rfam RNA family database. RNAlien

models are a starting point to construct models of comparable sensitivity and

specificity to manually curated ones from the Rfam database. RNAlien Tool and web

server are available at http://rna.tbi.univie.ac.at/rnalien/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw558

PMCID: PMC5041467

PMID: 27330139

105. J Chem Inf Model. 2016 Sep 26;56(9):1615-21. doi: 10.1021/acs.jcim.6b00397. Epub

2016 Aug 19.

Connection Map for Compounds (CMC): A Server for Combinatorial Drug Toxicity and

Efficacy Analysis.

Liu L(1), Tsompana M(2), Wang Y(3), Wu D(1), Zhu L(4,)(5,)(6), Zhu R(1).

Author information:

(1)Department of Bioinformatics, School of Life Sciences and Technology, Tongji

University , Shanghai 200092, People's Repubic of China. (2)Center of Excellence

in Bioinformatics and Life Sciences, the State University of New York at Buffalo

, Buffalo, New York 14203, United States. (3)Basic Medical College, Beijing

University of Chinese Medicine , Beijing 100029, People's Republic of China.

(4)Digestive Diseases and Nutrition Center, Department of Pediatrics, The State

University of New York at Buffalo , Buffalo, New York 14260, United States.

(5)Genome, Environment, and Microbiome Community of Excellence, The State

University of New York at Buffalo , Buffalo, New York 14214, United States.

(6)Institute of Digestive Diseases, Longhua Hospital, Shanghai University of

Traditional Chinese Medicine , Shanghai 200032, People's Republic of China.

Drug discovery and development is a costly and time-consuming process with a high

risk for failure resulting primarily from a drug's associated clinical safety and

efficacy potential. Identifying and eliminating inapt candidate drugs as early as

possible is an effective way for reducing unnecessary costs, but limited

analytical tools are currently available for this purpose. Recent growth in the

area of toxicogenomics and pharmacogenomics has provided with a vast amount of

drug expression microarray data. Web servers such as CMap and LTMap have used

this information to evaluate drug toxicity and mechanisms of action

independently; however, their wider applicability has been limited by the lack of

a combinatorial drug-safety type of analysis. Using available genome-wide drug

transcriptional expression profiles, we developed the first web server for

combinatorial evaluation of toxicity and efficacy of candidate drugs named

"Connection Map for Compounds" (CMC). Using CMC, researchers can initially

compare their query drug gene signatures with prebuilt gene profiles generated

from two large-scale toxicogenomics databases, and subsequently perform a drug

efficacy analysis for identification of known mechanisms of drug action or

generation of new predictions. CMC provides a novel approach for drug

repositioning and early evaluation in drug discovery with its unique combination

of toxicity and efficacy analyses, expansibility of data and algorithms, and

customization of reference gene profiles. CMC can be freely accessed at

http://cadd.tongji.edu.cn/webserver/CMCbp.jsp .

DOI: 10.1021/acs.jcim.6b00397

PMID: 27508329

106. PLoS One. 2016 Sep 23;11(9):e0163274. doi: 10.1371/journal.pone.0163274.

Sequence Based Prediction of Antioxidant Proteins Using a Classifier Selection

Strategy.

Zhang L(1), Zhang C(1,)(2), Gao R(1), Yang R(1), Song Q(3).

Author information:

(1)School of Control Science and Engineering, Shandong University, Jinan, China.

(2)School of Mechanical, Electrical and Information Engineering, Shandong

University at Weihai, China. (3)School of Electrical Engineering, University of

Jinan, Jinan, China.

Antioxidant proteins perform significant functions in maintaining

oxidation/antioxidation balance and have potential therapies for some diseases.

Accurate identification of antioxidant proteins could contribute to revealing

physiological processes of oxidation/antioxidation balance and developing novel

antioxidation-based drugs. In this study, an ensemble method is presented to

predict antioxidant proteins with hybrid features, incorporating SSI (Secondary

Structure Information), PSSM (Position Specific Scoring Matrix), RSA (Relative

Solvent Accessibility), and CTD (Composition, Transition, Distribution). The

prediction results of the ensemble predictor are determined by an average of

prediction results of multiple base classifiers. Based on a classifier selection

strategy, we obtain an optimal ensemble classifier composed of RF (Random

Forest), SMO (Sequential Minimal Optimization), NNA (Nearest Neighbor Algorithm),

and J48 with an accuracy of 0.925. A Relief combined with IFS (Incremental

Feature Selection) method is adopted to obtain optimal features from hybrid

features. With the optimal features, the ensemble method achieves improved

performance with a sensitivity of 0.95, a specificity of 0.93, an accuracy of

0.94, and an MCC (Matthew's Correlation Coefficient) of 0.880, far better than

the existing method. To evaluate the prediction performance objectively, the

proposed method is compared with existing methods on the same independent testing

dataset. Encouragingly, our method performs better than previous studies. In

addition, our method achieves more balanced performance with a sensitivity of

0.878 and a specificity of 0.860. These results suggest that the proposed

ensemble method can be a potential candidate for antioxidant protein prediction.

For public access, we develop a user-friendly web server for antioxidant protein

identification that is freely accessible at http://antioxidant.weka.cc.

DOI: 10.1371/journal.pone.0163274

PMCID: PMC5035026

PMID: 27662651

Conflict of interest statement: The authors have declared that no competing

interests exist.

107. PLoS One. 2016 Sep 22;11(9):e0163454. doi: 10.1371/journal.pone.0163454.

VfoldCPX Server: Predicting RNA-RNA Complex Structure and Stability.

Xu X(1), Chen SJ(1,)(2,)(3).

Author information:

(1)Department of Physics, University of Missouri, Columbia, MO 65211, United

States of America. (2)Department of Biochemistry, University of Missouri,

Columbia, MO 65211, United States of America. (3)MU Informatics Institute,

University of Missouri, Columbia, MO 65211, United States of America.

RNA-RNA interactions are essential for genomic RNA dimerization, mRNA splicing,

and many RNA-related gene expression and regulation processes. The prediction of

the structure and folding stability of RNA-RNA complexes is a problem of

significant biological importance and receives substantial interest in the

biological community. The VfoldCPX server provides a new web interface to predict

the two-dimensional (2D) structures of RNA-RNA complexes from the nucleotide

sequences. The VfoldCPX server has several novel advantages including the ability

to treat RNAs with tertiary contacts (crossing base pairs) such as loop-loop

kissing interactions and the use of physical loop entropy parameters. Based on a

partition function-based algorithm, the server enables prediction for structure

with and without tertiary contacts. Furthermore, the server outputs a set of

energetically stable structures, ranked by their stabilities. The results allow

users to gain extensive physical insights into RNA-RNA interactions and their

roles in RNA function. The web server is freely accessible at

"http://rna.physics.missouri.edu/vfoldCPX".

DOI: 10.1371/journal.pone.0163454

PMCID: PMC5033388

PMID: 27657918

Conflict of interest statement: The authors have declared that no competing

interests exist.

108. Biol Direct. 2016 Sep 21;11(1):48.

modPDZpep: a web resource for structure based analysis of human PDZ-mediated

interaction networks.

Sain N(1), Mohanty D(2).

Author information:

(1)Bioinformatics Center, National Institute of Immunology, Aruna Asaf Ali Marg,

New Delhi, 110067, India. (2)Bioinformatics Center, National Institute of

Immunology, Aruna Asaf Ali Marg, New Delhi, 110067, India. deb@nii.res.in.

BACKGROUND: PDZ domains recognize short sequence stretches usually present in

C-terminal of their interaction partners. Because of the involvement of PDZ

domains in many important biological processes, several attempts have been made

for developing bioinformatics tools for genome-wide identification of PDZ

interaction networks. Currently available tools for prediction of interaction

partners of PDZ domains utilize machine learning approach. Since, they have been

trained using experimental substrate specificity data for specific PDZ families,

their applicability is limited to PDZ families closely related to the training

set. These tools also do not allow analysis of PDZ-peptide interaction

interfaces.

RESULTS: We have used a structure based approach to develop modPDZpep, a program

to predict the interaction partners of human PDZ domains and analyze structural

details of PDZ interaction interfaces. modPDZpep predicts interaction partners by

using structural models of PDZ-peptide complexes and evaluating binding energy

scores using residue based statistical pair potentials. Since, it does not

require training using experimental data on peptide binding affinity, it can

predict substrates for diverse PDZ families. Because of the use of simple scoring

function for binding energy, it is also fast enough for genome scale structure

based analysis of PDZ interaction networks. Benchmarking using artificial as well

as real negative datasets indicates good predictive power with ROC-AUC values in

the range of 0.7 to 0.9 for a large number of human PDZ domains. Another novel

feature of modPDZpep is its ability to map novel PDZ mediated interactions in

human protein-protein interaction networks, either by utilizing available

experimental phage display data or by structure based predictions.

CONCLUSIONS: In summary, we have developed modPDZpep, a web-server for structure

based analysis of human PDZ domains. It is freely available at

http://www.nii.ac.in/modPDZpep.html or http://202.54.226.235/modPDZpep.html .

REVIEWERS: This article was reviewed by Michael Gromiha and Zoltán Gáspári.

DOI: 10.1186/s13062-016-0151-4

PMCID: PMC5031328

PMID: 27655048

109. Bioinformatics. 2016 Sep 15;32(18):2850-2. doi: 10.1093/bioinformatics/btw238.

Epub 2016 Jun 6.

MetalPredator: a web server to predict iron-sulfur cluster binding proteomes.

Valasatava Y(1), Rosato A(2), Banci L(2), Andreini C(2).

Author information:

(1)Magnetic Resonance Center (CERM). (2)Magnetic Resonance Center (CERM)

Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy.

MOTIVATION: The prediction of the iron-sulfur proteome is highly desirable for

biomedical and biological research but a freely available tool to predict

iron-sulfur proteins has not been developed yet.

RESULTS: We developed a web server to predict iron-sulfur proteins from protein

sequence(s). This tool, called MetalPredator, is able to process complete

proteomes rapidly with high recall and precision.

AVAILABILITY AND IMPLEMENTATION: The web server is freely available at:

http://metalweb.cerm.unifi.it/tools/metalpredator/

CONTACT: andreini@cerm.unifi.it

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw238

PMID: 27273670

110. Bioinformatics. 2016 Sep 15;32(18):2853-5. doi: 10.1093/bioinformatics/btw315.

Epub 2016 Jun 3.

DBSI server: DNA binding site identifier.

Sukumar S(1), Zhu X(2), Ericksen SS(3), Mitchell JC(4).

Author information:

(1)Department of Biochemistry, University of Wisconsin-Madison, Madison, WI

53706, USA. (2)School of Life Sciences, Anhui University, Hefei, Anhui Province

230601, China. (3)Small Molecule Screening Facility. (4)Department of

Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA Department

of Mathematics, University of Wisconsin-Madison, Madison, WI 53706, USA.

: Protein-nucleic acid interactions are among the most important intermolecular

interactions in the regulation of cellular events. Identifying residues involved

in these interactions from protein structure alone is an important challenge.

Here we introduce the webserver interface to DNA Binding Site Identifier (DBSI),

a powerful structure-based SVM model for the prediction and visualization of DNA

binding sites on protein structures. DBSI has been shown to be a top-performing

model to predict DNA binding sites on the surface of a protein or peptide and

shows promise in predicting RNA binding sites.AVAILABILITY AND IMPLEMENTATION:

Server is available at http://dbsi.mitchell-lab.org

CONTACT: jcmitchell@wisc.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw315

PMID: 27259543

111. PLoS One. 2016 Sep 15;11(9):e0162442. doi: 10.1371/journal.pone.0162442.

eCollection 2016.

MetaStorm: A Public Resource for Customizable Metagenomics Annotation.

Arango-Argoty G(1), Singh G(2), Heath LS(1), Pruden A(3), Xiao W(4), Zhang L(1).

Author information:

(1)Department of Computer Science, Virginia Tech, Blacksburg, Virginia, United

States of America. (2)Department of Civil Engineering, Indian Institute of

Technology Roorkee, Roorkee, Uttarakhand, India. (3)Department of Civil and

Environmental Engineering, Virginia Tech, Blacksburg, Virginia, United States of

America. (4)Department of Microbiology and Immunology, Temple University School

of Medicine, Philadelphia, United States of America.

Metagenomics is a trending research area, calling for the need to analyze large

quantities of data generated from next generation DNA sequencing technologies.

The need to store, retrieve, analyze, share, and visualize such data challenges

current online computational systems. Interpretation and annotation of specific

information is especially a challenge for metagenomic data sets derived from

environmental samples, because current annotation systems only offer broad

classification of microbial diversity and function. Moreover, existing resources

are not configured to readily address common questions relevant to environmental

systems. Here we developed a new online user-friendly metagenomic analysis server

called MetaStorm (http://bench.cs.vt.edu/MetaStorm/), which facilitates

customization of computational analysis for metagenomic data sets. Users can

upload their own reference databases to tailor the metagenomics annotation to

focus on various taxonomic and functional gene markers of interest. MetaStorm

offers two major analysis pipelines: an assembly-based annotation pipeline and

the standard read annotation pipeline used by existing web servers. These

pipelines can be selected individually or together. Overall, MetaStorm provides

enhanced interactive visualization to allow researchers to explore and manipulate

taxonomy and functional annotation at various levels of resolution.

DOI: 10.1371/journal.pone.0162442

PMCID: PMC5025195

PMID: 27632579

Conflict of interest statement: The authors have declared that no competing

interests exist.

112. BMC Genomics. 2016 Sep 9;17(1):722. doi: 10.1186/s12864-016-3057-8.

Mergeomics: a web server for identifying pathological pathways, networks, and key

regulators via multidimensional data integration.

Arneson D(1), Bhattacharya A(1), Shu L(1), Mäkinen VP(2,)(3,)(4), Yang X(5).

Author information:

(1)Department of Integrative Biology and Physiology, University of California,

Los Angeles, CA, 90095, USA. (2)South Australian Health and Medical Research

Institute, Adelaide, Australia. (3)School of Biological Sciences, University of

Adelaide, Adelaide, Australia. (4)Institute of Health Sciences, University of

Oulu, Oulu, Finland. (5)Department of Integrative Biology and Physiology,

University of California, Los Angeles, CA, 90095, USA. xyang123@ucla.edu.

BACKGROUND: Human diseases are commonly the result of multidimensional changes at

molecular, cellular, and systemic levels. Recent advances in genomic technologies

have enabled an outpour of omics datasets that capture these changes. However,

separate analyses of these various data only provide fragmented understanding and

do not capture the holistic view of disease mechanisms. To meet the urgent needs

for tools that effectively integrate multiple types of omics data to derive

biological insights, we have developed Mergeomics, a computational pipeline that

integrates multidimensional disease association data with functional genomics and

molecular networks to retrieve biological pathways, gene networks, and central

regulators critical for disease development.

RESULTS: To make the Mergeomics pipeline available to a wider research community,

we have implemented an online, user-friendly web server ( http://mergeomics.

RESEARCH: idre.ucla.edu/ ). The web server features a modular implementation of

the Mergeomics pipeline with detailed tutorials. Additionally, it provides

curated genomic resources including tissue-specific expression quantitative trait

loci, ENCODE functional annotations, biological pathways, and molecular networks,

and offers interactive visualization of analytical results. Multiple

computational tools including Marker Dependency Filtering (MDF), Marker Set

Enrichment Analysis (MSEA), Meta-MSEA, and Weighted Key Driver Analysis (wKDA)

can be used separately or in flexible combinations. User-defined summary-level

genomic association datasets (e.g., genetic, transcriptomic, epigenomic) related

to a particular disease or phenotype can be uploaded and computed real-time to

yield biologically interpretable results, which can be viewed online and

downloaded for later use.

CONCLUSIONS: Our Mergeomics web server offers researchers flexible and

user-friendly tools to facilitate integration of multidimensional data into

holistic views of disease mechanisms in the form of tissue-specific key

regulators, biological pathways, and gene networks.

DOI: 10.1186/s12864-016-3057-8

PMCID: PMC5016927

PMID: 27612452

113. Curr Protoc Bioinformatics. 2016 Sep 7;55:12.14.1-12.14.18. doi: 10.1002/cpbi.12.

DIANA-TarBase and DIANA Suite Tools: Studying Experimentally Supported microRNA

Targets.

Paraskevopoulou MD(1,)(2), Vlachos IS(1,)(2), Hatzigeorgiou AG(1,)(2).

Author information:

(1)DIANA-Lab, Department of Electrical & Computer Engineering, University of

Thessaly, Volos, Greece. (2)Hellenic Pasteur Institute, Athens, Greece.

microRNAs (miRNAs) are short non-coding RNAs (∼22 nts) present in animals,

plants, and viruses. They are considered central post-transcriptional regulators

of gene expression and are key components in a great number of physiological and

pathological conditions. The accurate characterization of their targets is

considered essential to a series of applications and basic or applied research

settings. DIANA-TarBase (http://www.microrna.gr/tarbase) was initially launched

in 2006. It is a reference repository indexing experimentally derived miRNA-gene

interactions in different cell types, tissues, and conditions across numerous

species. This unit focuses on the study of experimentally supported miRNA-gene

interactions, as well as their functional interpretation through the use of

available tools in the DIANA suite (http://www.microrna.gr). The proposed

use-case scenarios are presented in protocols, describing how to utilize the

DIANA-TarBase database and DIANA-microT-CDS server and perform miRNA-targeted

pathway analysis with DIANA-miRPath-v3. All analyses are directly invoked or

initiated from DIANA-TarBase. © 2016 by John Wiley & Sons, Inc.

Copyright © 2016 John Wiley & Sons, Inc.

DOI: 10.1002/cpbi.12

PMID: 27603020

114. J Theor Biol. 2016 Sep 7;404:285-94. doi: 10.1016/j.jtbi.2016.06.013. Epub 2016

Jun 11.

A computational approach for prediction of donor splice sites with improved

accuracy.

Meher PK(1), Sahu TK(2), Rao AR(3), Wahi SD(4).

Author information:

(1)ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012,

India. Electronic address: meherprabin@yahoo.com. (2)ICAR-Indian Agricultural

Statistics Research Institute, New Delhi 110012, India. Electronic address:

tanmayabioinfo@gmail.com. (3)ICAR-Indian Agricultural Statistics Research

Institute, New Delhi 110012, India. Electronic address: rao.cshl.work@gmail.com.

(4)ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012,

India. Electronic address: sdwahi@iasri.res.in.

Identification of splice sites is important due to their key role in predicting

the exon-intron structure of protein coding genes. Though several approaches have

been developed for the prediction of splice sites, further improvement in the

prediction accuracy will help predict gene structure more accurately. This paper

presents a computational approach for prediction of donor splice sites with

higher accuracy. In this approach, true and false splice sites were first encoded

into numeric vectors and then used as input in artificial neural network (ANN),

support vector machine (SVM) and random forest (RF) for prediction. ANN and SVM

were found to perform equally and better than RF, while tested on HS3D and NN269

datasets. Further, the performance of ANN, SVM and RF were analyzed by using an

independent test set of 50 genes and found that the prediction accuracy of ANN

was higher than that of SVM and RF. All the predictors achieved higher accuracy

while compared with the existing methods like NNsplice, MEM, MDD, WMM, MM1,

FSPLICE, GeneID and ASSP, using the independent test set. We have also developed

an online prediction server (PreDOSS) available at

http://cabgrid.res.in:8080/predoss, for prediction of donor splice sites using

the proposed approach.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2016.06.013

PMID: 27302911

115. Biochim Biophys Acta. 2016 Sep;1864(9):1104-9. doi: 10.1016/j.bbapap.2016.06.001.

Epub 2016 Jun 2.

Prediction of change in protein unfolding rates upon point mutations in two state

proteins.

Chaudhary P(1), Naganathan AN(1), Gromiha MM(2).

Author information:

(1)Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences,

Indian Institute of Technology Madras, Chennai 600 036, Tamil Nadu, India.

(2)Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences,

Indian Institute of Technology Madras, Chennai 600 036, Tamil Nadu, India.

Electronic address: gromiha@iitm.ac.in.

Studies on protein unfolding rates are limited and challenging due to the

complexity of unfolding mechanism and the larger dynamic range of the

experimental data. Though attempts have been made to predict unfolding rates

using protein sequence-structure information there is no available method for

predicting the unfolding rates of proteins upon specific point mutations. In this

work, we have systematically analyzed a set of 790 single mutants and developed a

robust method for predicting protein unfolding rates upon mutations (Δlnku) in

two-state proteins by combining amino acid properties and knowledge-based

classification of mutants with multiple linear regression technique. We obtain a

mean absolute error (MAE) of 0.79/s and a Pearson correlation coefficient (PCC)

of 0.71 between predicted unfolding rates and experimental observations using

jack-knife test. We have developed a web server for predicting protein unfolding

rates upon mutation and it is freely available at

https://www.iitm.ac.in/bioinfo/proteinunfolding/unfoldingrace.html. Prominent

features that determine unfolding kinetics as well as plausible reasons for the

observed outliers are also discussed.

Copyright © 2016 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.bbapap.2016.06.001

PMID: 27264959

116. Bioinformatics. 2016 Sep 1;32(17):2710-2. doi: 10.1093/bioinformatics/btw301.

Epub 2016 May 13.

PRESS: PRotEin S-Sulfenylation server.

Sakka M(1), Tzortzis G(2), Mantzaris MD(3), Bekas N(1), Kellici TF(1), Likas

A(2), Galaris D(3), Gerothanassis IP(1), Tzakos AG(1).

Author information:

(1)Department of Chemistry. (2)Department of Computer Science and Engineering.

(3)Laboratory of Biological Chemistry Faculty of Medicine, University of

Ioannina, Ioannina, Greece.

MOTIVATION: Transient S-sulfenylation of cysteine thiols mediated by reactive

oxygen species plays a critical role in pathology, physiology and cell signaling.

Therefore, discovery of new S-sulfenylated sites in proteins is of great

importance towards understanding how protein function is regulated upon redox

conditions.

RESULTS: We developed PRESS (PRotEin S-Sulfenylation) web server, a server which

can effectively predict the cysteine thiols of a protein that could undergo

S-sulfenylation under redox conditions. We envisage that this server will boost

and facilitate the discovery of new and currently unknown functions of proteins

triggered upon redox conditions, signal regulation and transduction, thus

uncovering the role of S-sulfenylation in human health and disease.

AVAILABILITY AND IMPLEMENTATION: The PRESS web server is freely available at

http://press-sulfenylation.cse.uoi.gr/

CONTACTS: agtzakos@gmail.com or gtzortzi@cs.uoi.gr

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw301

PMID: 27187205

117. J Biomol Struct Dyn. 2016 Sep;34(9):1946-61. doi: 10.1080/07391102.2015.1095116.

Epub 2015 Oct 29.

Identification of protein-protein binding sites by incorporating the

physicochemical properties and stationary wavelet transforms into pseudo amino

acid composition.

Jia J(1), Liu Z(1), Xiao X(1,)(2), Liu B(1), Chou KC(3,)(2).

Author information:

(1)a Computer Department , Jing-De-Zhen Ceramic Institute , Jing-De-Zhen 333403 ,

China. (2)c Gordon Life Science Institute , Boston , MA 02478 , USA. (3)b Center

of Excellence in Genomic Medicine Research (CEGMR) , King Abdulaziz University ,

Jeddah 21589 , Saudi Arabia.

With the explosive growth of protein sequences entering into protein data banks

in the post-genomic era, it is highly demanded to develop automated methods for

rapidly and effectively identifying the protein-protein binding sites (PPBSs)

based on the sequence information alone. To address this problem, we proposed a

predictor called iPPBS-PseAAC, in which each amino acid residue site of the

proteins concerned was treated as a 15-tuple peptide segment generated by sliding

a window along the protein chains with its center aligned with the target

residue. The working peptide segment is further formulated by a general form of

pseudo amino acid composition via the following procedures: (1) it is converted

into a numerical series via the physicochemical properties of amino acids; (2)

the numerical series is subsequently converted into a 20-D feature vector by

means of the stationary wavelet transform technique. Formed by many individual

"Random Forest" classifiers, the operation engine to run prediction is a

two-layer ensemble classifier, with the 1st-layer voting out the best training

data-set from many bootstrap systems and the 2nd-layer voting out the most

relevant one from seven physicochemical properties. Cross-validation tests

indicate that the new predictor is very promising, meaning that many important

key features, which are deeply hidden in complicated protein sequences, can be

extracted via the wavelets transform approach, quite consistent with the facts

that many important biological functions of proteins can be elucidated with their

low-frequency internal motions. The web server of iPPBS-PseAAC is accessible at

http://www.jci-bioinfo.cn/iPPBS-PseAAC , by which users can easily acquire their

desired results without the need to follow the complicated mathematical equations

involved.

DOI: 10.1080/07391102.2015.1095116

PMID: 26375780 [Indexed for MEDLINE]

118. J Radiol Prot. 2016 Sep;36(3):561-578. Epub 2016 Jul 27.

InterCardioRisk: a novel online tool for estimating doses of ionising radiation

to occupationally-exposed medical staff and their associated health risks.

Moriña D(1), Grellier J, Carnicer A, Pernot E, Ryckx N, Cardis E.

Author information:

(1)ISGlobal, Centre for Research in Environmental Epidemiology (CREAL),

Barcelona, Spain. Universitat Pompeu Fabra (UPF), Barcelona, Spain. CIBER

Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain. Unit of Infections

and Cancer (UNIC), Cancer Epidemiology Research Program (CERP), Catalan Institute

of Oncology (ICO)-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain. Grups de

Recerca d'Àfrica i Amèrica Llatines (GRAAL)-Unitat de Bioestadística, Facultat de

Medicina, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain.

Those working in interventional cardiology and related medical procedures are

potentially subject to considerable exposure to x-rays. Two types of tissue of

particular concern that may receive considerable doses during such procedures are

the lens of the eye and the brain. Ocular radiation exposure results in lens

changes that, with time, may progress to partial or total lens opacification

(cataracts). In the early stages, such opacities do not result in visual

disability; the severity of such changes tends to increase progressively with

dose and time until vision is impaired and cataract surgery is required.

Scattered radiation doses to the eye lens of an interventional cardiologist in

typical working conditions can exceed 34 μGy min(-1) in high-dose fluoroscopy

modes and 3 μGy per image during image acquisition (instantaneous rate values)

when radiation protection tools are not used. A causal relation between exposure

to ionising radiation and increased risk of brain and central nervous system

tumours has been shown in a number of studies. Although absorbed doses to the

brain in interventional cardiology procedures are lower than those to the eye

lens by a factor between 3.40 and 8.08 according to our simulations, doses to

both tissues are among the highest occupational radiation doses documented for

medical staff whose work involves exposures to x-rays. We present

InterCardioRisk, a tool featuring an easy-to-use web interface that provides a

general estimation of both cumulated absorbed doses experienced by medical staff

exposed in the interventional cardiology setting and their estimated associated

health risks. The tool is available at http://intercardiorisk.creal.cat.

DOI: 10.1088/0952-4746/36/3/561

PMID: 27460876

119. Methods. 2016 Sep 1;107:34-41. doi: 10.1016/j.ymeth.2016.03.013. Epub 2016 Mar

23.

tRNAmodpred: A computational method for predicting posttranscriptional

modifications in tRNAs.

Machnicka MA(1), Dunin-Horkawicz S(1), de Crécy-Lagard V(2), Bujnicki JM(3).

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland.

(2)Department of Microbiology and Cell Science, University Florida, Gainesville,

FL, USA. (3)Laboratory of Bioinformatics and Protein Engineering, International

Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw,

Poland; Institute of Molecular Biology and Biotechnology, Adam Mickiewicz

University, Umultowska 89, PL-61-614 Poznan, Poland. Electronic address:

iamb@genesilico.pl.

tRNA molecules contain numerous chemically altered nucleosides, which are formed

by enzymatic modification of the primary transcripts during the complex tRNA

maturation process. Some of the modifications are introduced by single reactions,

while other require complex series of reactions carried out by several different

enzymes. The location and distribution of various types of modifications vary

greatly between different tRNA molecules, organisms and organelles. We have

developed a computational method tRNAmodpred, for predicting modifications in

tRNA sequences. Briefly, our method takes as an input one or more unmodified tRNA

sequences and a set of protein sequences corresponding to a proteome of a cell.

Subsequently it identifies homologs of known tRNA modification enzymes in the

proteome, predicts tRNA modification activities and maps them onto known pathways

of RNA modification from the MODOMICS database. Thereby, theoretically possible

modification pathways are identified, and products of these modification

reactions are proposed for query tRNAs. This method allows for predicting

modification patterns for newly sequenced genomes as well as for checking

tentative modification status of tRNAs from one species treated with enzymes from

another source, e.g. to predict the possible modifications of eukaryotic tRNAs

expressed in bacteria. tRNAmodpred is freely available as a web server at

http://genesilico.pl/trnamodpred/.

Copyright © 2016 The Authors. Published by Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ymeth.2016.03.013

PMCID: PMC5014562 [Available on 2017-09-01]

PMID: 27016142

120. Probiotics Antimicrob Proteins. 2016 Sep;8(3):141-9. doi:

10.1007/s12602-016-9215-0.

Development of Antimicrobial Peptide Prediction Tool for Aquaculture Industries.

Gautam A(1), Sharma A(1), Jaiswal S(2), Fatma S(2), Arora V(2), Iquebal MA(2),

Nandi S(3), Sundaray JK(3), Jayasankar P(3), Rai A(2), Kumar D(4).

Author information:

(1)Jaypee University of Information Technology, Solan, Himachal Pradesh, India.

(2)Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics

Research Institute, New Delhi, India. (3)Division of Fish Genetics and

Biotechnology, ICAR-Central Institute of Freshwater Aquaculture, Bhubaneswar,

Odisha, India. (4)Centre for Agricultural Bioinformatics, ICAR-Indian

Agricultural Statistics Research Institute, New Delhi, India.

dineshkumarbhu@gmail.com.

Microbial diseases in fish, plant, animal and human are rising constantly; thus,

discovery of their antidote is imperative. The use of antibiotic in aquaculture

further compounds the problem by development of resistance and consequent

consumer health risk by bio-magnification. Antimicrobial peptides (AMPs) have

been highly promising as natural alternative to chemical antibiotics. Though AMPs

are molecules of innate immune defense of all advance eukaryotic organisms, fish

being heavily dependent on their innate immune defense has been a good source of

AMPs with much wider applicability. Machine learning-based prediction method

using wet laboratory-validated fish AMP can accelerate the AMP discovery using

available fish genomic and proteomic data. Earlier AMP prediction servers are

based on multi-phyla/species data, and we report here the world's first AMP

prediction server in fishes. It is freely accessible at

http://webapp.cabgrid.res.in/fishamp/ . A total of 151 AMPs related to fish

collected from various databases and published literature were taken for this

study. For model development and prediction, N-terminus residues, C-terminus

residues and full sequences were considered. Best models were with kernels

polynomial-2, linear and radial basis function with accuracy of 97, 99 and 97 %,

respectively. We found that performance of support vector machine-based models is

superior to artificial neural network. This in silico approach can drastically

reduce the time and cost of AMP discovery. This accelerated discovery of lead AMP

molecules having potential wider applications in diverse area like fish and human

health as substitute of antibiotics, immunomodulator, antitumor, vaccine adjuvant

and inactivator, and also for packaged food can be of much importance for

industries.

DOI: 10.1007/s12602-016-9215-0

PMID: 27141850

121. Sci Rep. 2016 Sep 1;6:32333. doi: 10.1038/srep32333.

dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein

sequence and rank aggregation.

Chen J(1), Long R(1), Wang XL(1,)(2), Liu B(1,)(2,)(3), Chou KC(3,)(4).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong 518055, China. (2)Key Laboratory of

Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, Guangdong 518055, China. (3)Gordon Life Science

Institute, Boston, MA 02478, USA. (4)Center of Excellence in Genomic Medicine

Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia.

Protein remote homology detection is an important task in computational

proteomics. Some computational methods have been proposed, which detect remote

homology proteins based on different features and algorithms. As noted in

previous studies, their predictive results are complementary to each other.

Therefore, it is intriguing to explore whether these methods can be combined into

one package so as to further enhance the performance power and application

convenience. In view of this, we introduced a protein representation called

profile-based pseudo protein sequence to extract the evolutionary information

from the relevant profiles. Based on the concept of pseudo proteins, a new

predictor, called "dRHP-PseRA", was developed by combining four state-of-the-art

predictors (PSI-BLAST, HHblits, Hmmer, and Coma) via the rank aggregation

approach. Cross-validation tests on a SCOP benchmark dataset have demonstrated

that the new predictor has remarkably outperformed any of the existing methods

for the same purpose on ROC50 scores. Accordingly, it is anticipated that

dRHP-PseRA holds very high potential to become a useful high throughput tool for

detecting remote homology proteins. For the convenience of most experimental

scientists, a web-server for dRHP-PseRA has been established at

http://bioinformatics.hitsz.edu.cn/dRHP-PseRA/.

DOI: 10.1038/srep32333

PMCID: PMC5007510

PMID: 27581095

122. BMC Genomics. 2016 Aug 26;17:681. doi: 10.1186/s12864-016-3028-0.

G23D: Online tool for mapping and visualization of genomic variants on 3D protein

structures.

Solomon O(1,)(2), Kunik V(1), Simon A(3), Kol N(1), Barel O(1), Lev A(3),

Amariglio N(1,)(2), Somech R(4,)(5), Rechavi G(1,)(5), Eyal E(6).

Author information:

(1)Cancer Research Center, Sheba Medical Center, Ramat-Gan, Israel. (2)The Mina

and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan,

Israel. (3)Pediatric Immunology Service, Jeffrey Modell Foundation, Sheba Medical

Center, Ramat-Gan, Israel. (4)Edmond and Lily Safra Children's Hospital, Sheba

Medical Center, Ramat-Gan, Israel. (5)Sackler School of Medicine, Tel Aviv

University, Tel Aviv, Israel. (6)Cancer Research Center, Sheba Medical Center,

Ramat-Gan, Israel. eran.eyal@sheba.health.gov.il.

BACKGROUND: Evaluation of the possible implications of genomic variants is an

increasingly important task in the current high throughput sequencing era.

Structural information however is still not routinely exploited during this

evaluation process. The main reasons can be attributed to the partial structural

coverage of the human proteome and the lack of tools which conveniently convert

genomic positions, which are the frequent output of genomic pipelines, to

proteins and structure coordinates.

RESULTS: We present G23D, a tool for conversion of human genomic coordinates to

protein coordinates and protein structures. G23D allows mapping of genomic

positions/variants on evolutionary related (and not only identical) protein three

dimensional (3D) structures as well as on theoretical models. By doing so it

significantly extends the space of variants for which structural insight is

feasible. To facilitate interpretation of the variant consequence, pathogenic

variants, functional sites and polymorphism sites are displayed on protein

sequence and structure diagrams alongside the input variants. G23D also provides

modeling of the mutant structure, analysis of intra-protein contacts and instant

access to functional predictions and predictions of thermo-stability changes.

G23D is available at http://www.sheba-cancer.org.il/G23D .

CONCLUSIONS: G23D extends the fraction of variants for which structural analysis

is applicable and provides better and faster accessibility for structural data to

biologists and geneticists who routinely work with genomic information.

DOI: 10.1186/s12864-016-3028-0

PMCID: PMC5002099

PMID: 27565432

123. J Cheminform. 2016 Aug 26;8(1):42. doi: 10.1186/s13321-016-0155-1. eCollection

2016.

Molmil: a molecular viewer for the PDB and beyond.

Bekker GJ(1), Nakamura H(2), Kinjo AR(2).

Author information:

(1)Laboratory of Protein Informatics, Institute for Protein Research, Osaka

University, 3-2 Yamadaoka, Suita, Osaka 565-0871 Japan ; Graduate School of

Frontier Biosciences, Osaka University, Suita, Osaka 565-0871 Japan.

(2)Laboratory of Protein Informatics, Institute for Protein Research, Osaka

University, 3-2 Yamadaoka, Suita, Osaka 565-0871 Japan.

We have developed a new platform-independent web-based molecular viewer using

JavaScript and WebGL. The molecular viewer, Molmil, has been integrated into

several services offered by Protein Data Bank Japan and can be easily extended

with new functionality by third party developers. Furthermore, the viewer can be

used to load files in various formats from the user's local hard drive without

uploading the data to a server. Molmil is available for all platforms supporting

WebGL (e.g. Windows, Linux, iOS, Android) from http://gjbekker.github.io/molmil/.

The source code is available at http://github.com/gjbekker/molmil under the

LGPLv3 licence.

DOI: 10.1186/s13321-016-0155-1

PMCID: PMC5002144

PMID: 27570544

124. PLoS One. 2016 Aug 25;11(8):e0160645. doi: 10.1371/journal.pone.0160645.

eCollection 2016.

aPPRove: An HMM-Based Method for Accurate Prediction of RNA-Pentatricopeptide

Repeat Protein Binding Events.

Harrison T(1), Ruiz J(2), Sloan DB(3), Ben-Hur A(1), Boucher C(2).

Author information:

(1)Department of Computer Science, Colorado State University, Fort Collins, CO,

80523, United States of America. (2)Department of Computer and Information

Science and Engineering, University of Florida, Gainesville, FL, 32611, United

States of America. (3)Department of Biology, Colorado State University, Fort

Collins, CO, 80523, United States of America.

Pentatricopeptide repeat containing proteins (PPRs) bind to RNA transcripts

originating from mitochondria and plastids. There are two classes of PPR

proteins. The [Formula: see text] class contains tandem [Formula: see text]-type

motif sequences, and the [Formula: see text] class contains alternating [Formula:

see text], [Formula: see text] and [Formula: see text] type sequences. In this

paper, we describe a novel tool that predicts PPR-RNA interaction; specifically,

our method, which we call aPPRove, determines where and how a [Formula: see

text]-class PPR protein will bind to RNA when given a PPR and one or more RNA

transcripts by using a combinatorial binding code for site specificity proposed

by Barkan et al. Our results demonstrate that aPPRove successfully locates how

and where a PPR protein belonging to the [Formula: see text] class can bind to

RNA. For each binding event it outputs the binding site, the

amino-acid-nucleotide interaction, and its statistical significance. Furthermore,

we show that our method can be used to predict binding events for [Formula: see

text]-class proteins using a known edit site and the statistical significance of

aligning the PPR protein to that site. In particular, we use our method to make a

conjecture regarding an interaction between CLB19 and the second intronic region

of ycf3. The aPPRove web server can be found at www.cs.colostate.edu/~approve.

DOI: 10.1371/journal.pone.0160645

PMCID: PMC4999063

PMID: 27560805

Conflict of interest statement: The authors have declared that no competing

interests exist.

125. Brief Bioinform. 2016 Aug 19. pii: bbw071. [Epub ahead of print]

A protein network descriptor server and its use in studying protein, disease,

metabolic and drug targeted networks.

Zhang P, Tao L, Zeng X, Qin C, Chen S, Zhu F, Li Z, Jiang Y, Chen W, Chen YZ.

The genetic, proteomic, disease and pharmacological studies have generated rich

data in protein interaction, disease regulation and drug activities useful for

systems-level study of the biological, disease and drug therapeutic processes.

These studies are facilitated by the established and the emerging computational

methods. More recently, the network descriptors developed in other disciplines

have become more increasingly used for studying the protein-protein, gene

regulation, metabolic, disease networks. There is an inadequate coverage of these

useful network features in the public web servers. We therefore introduced upto

313 literature-reported network descriptors in PROFEAT web server, for describing

the topological, connectivity and complexity characteristics of undirected

unweighted (uniform binding constants and molecular levels), undirected

edge-weighted (varying binding constants), undirected node-weighted (varying

molecular levels), undirected edge-node-weighted (varying binding constants and

molecular levels) and directed unweighted (oriented process) networks. The

usefulness of the PROFEAT computed network descriptors is illustrated by their

literature-reported applications in studying the protein-protein, gene

regulatory, gene co-expression, protein-drug and metabolic networks. PROFEAT is

accessible free of charge at

http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bib/bbw071

PMID: 27542402

126. Sci Rep. 2016 Aug 18;6:32153. doi: 10.1038/srep32153.

GalaxyRefineComplex: Refinement of protein-protein complex model structures

driven by interface repacking.

Heo L(1), Lee H(1), Seok C(1).

Author information:

(1)Department of Chemistry, Seoul National University, Seoul 08826, Republic of

Korea.

Protein-protein docking methods have been widely used to gain an atomic-level

understanding of protein interactions. However, docking methods that employ

low-resolution energy functions are popular because of computational efficiency.

Low-resolution docking tends to generate protein complex structures that are not

fully optimized. GalaxyRefineComplex takes such low-resolution docking structures

and refines them to improve model accuracy in terms of both interface contact and

inter-protein orientation. This refinement method allows flexibility at the

protein interface and in the overall docking structure to capture conformational

changes that occur upon binding. Symmetric refinement is also provided for

symmetric homo-complexes. This method was validated by refining models produced

by available docking programs, including ZDOCK and M-ZDOCK, and was successfully

applied to CAPRI targets in a blind fashion. An example of using the refinement

method with an existing docking method for ligand binding mode prediction of a

drug target is also presented. A web server that implements the method is freely

available at http://galaxy.seoklab.org/refinecomplex.

DOI: 10.1038/srep32153

PMCID: PMC4989233

PMID: 27535582

127. Bioinformatics. 2016 Aug 15;32(16):2548-50. doi: 10.1093/bioinformatics/btw208.

Epub 2016 Apr 19.

OREMPRO web server: orientation and assessment of atomistic and coarse-grained

structures of membrane proteins.

Postic G(1), Ghouzam Y(1), Gelly JC(1).

Author information:

(1)Inserm U1134, Paris, France Univ. Paris Diderot, Sorbonne Paris Cité, UMR\_S

1134, Paris, France Institut National De La Transfusion Sanguine, Paris, France

Laboratory of Excellence GR-Ex, Paris, France.

: The experimental determination of membrane protein orientation within the lipid

bilayer is extremely challenging, such that computational methods are most often

the only solution. Moreover, obtaining all-atom 3D structures of membrane

proteins is also technically difficult, and many of the available data are either

experimental low-resolution structures or theoretical models, whose structural

quality needs to be evaluated. Here, to address these two crucial problems, we

propose OREMPRO, a web server capable of both (i) positioning α-helical and

β-sheet transmembrane domains in the lipid bilayer and (ii) assessing their

structural quality. Most importantly, OREMPRO uses the sole alpha carbon

coordinates, which makes it the only web server compatible with both high and low

structural resolutions. Finally, OREMPRO is also interesting in its ability to

process coarse-grained protein models, by using coordinates of backbone beads in

place of alpha carbons.AVAILABILITY AND IMPLEMENTATION:

http://www.dsimb.inserm.fr/OREMPRO/ CONTACT: :

guillaume.postic@univ-paris-diderot.fr or

jean-christophe.gelly@univ-paris-diderot.fr

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw208

PMID: 27153644

128. Bioinformatics. 2016 Aug 15;32(16):2545-7. doi: 10.1093/bioinformatics/btw200.

Epub 2016 Apr 19.

Mollack: a web server for the automated creation of conformational ensembles for

intrinsically disordered proteins.

Ziegler Z(1), Schmidt M(2), Gurry T(3), Burger V(4), Stultz CM(5).

Author information:

(1)Cornell University, Ithaca, NY 14850, USA Research Laboratory of Electronics,

Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA. (2)Research

Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA

02139-4307, USA Department of Electrical Engineering and Computer Science,

Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA. (3)Research

Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA

02139-4307, USA Computational and Systems Biology Initiative, Massachusetts

Institute of Technology, Cambridge, MA 02139-4307, USA. (4)Research Laboratory of

Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307,

USA. (5)Research Laboratory of Electronics, Massachusetts Institute of

Technology, Cambridge, MA 02139-4307, USA Department of Electrical Engineering

and Computer Science, Massachusetts Institute of Technology, Cambridge, MA

02139-4307, USA Computational and Systems Biology Initiative, Massachusetts

Institute of Technology, Cambridge, MA 02139-4307, USA The Institute for Medical

Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA

02139-4307, USA.

Intrinsically disordered proteins (IDPs) play central roles in many biological

processes. Consequently, an accurate description of the disordered state is an

important step towards a comprehensive understanding of a number of important

biological functions. In this work we describe a new web server, Mollack, for the

automated construction of unfolded ensembles that uses both experimental and

molecular simulation data to construct models for the unfolded state. An

important aspect of the method is that it calculates a quantitative estimate of

the uncertainty in the constructed ensemble, thereby providing an objective

measure of the quality of the final model. Overall, Mollack facilitates

structure-function studies of disordered proteins.AVAILABILITY AND

IMPLEMENTATION: http://cmstultz-mollack.mit.edu

CONTACT: cmstultz@mit.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw200

PMID: 27153636

129. Bioinformatics. 2016 Aug 15;32(16):2542-4. doi: 10.1093/bioinformatics/btw192.

Epub 2016 Apr 10.

INPS-MD: a web server to predict stability of protein variants from sequence and

structure.

Savojardo C(1), Fariselli P(2), Martelli PL(1), Casadio R(3).

Author information:

(1)Biocomputing Group, CIG, Interdepartmental Center «Luigi Galvani» for

Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of

Bologna, Italy. (2)BCA, University of Padova, Italy. (3)Biocomputing Group, CIG,

Interdepartmental Center «Luigi Galvani» for Integrated Studies of

Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Italy

Interdepartmental Center «Giorgio Prodi» for Cancer Research, University of

Bologna, Italy.

MOTIVATION: Protein function depends on its structural stability. The effects of

single point variations on protein stability can elucidate the molecular

mechanisms of human diseases and help in developing new drugs. Recently, we

introduced INPS, a method suited to predict the effect of variations on protein

stability from protein sequence and whose performance is competitive with the

available state-of-the-art tools.

RESULTS: In this article, we describe INPS-MD (Impact of Non synonymous

variations on Protein Stability-Multi-Dimension), a web server for the prediction

of protein stability changes upon single point variation from protein sequence

and/or structure. Here, we complement INPS with a new predictor (INPS3D) that

exploits features derived from protein 3D structure. INPS3D scores with Pearson's

correlation to experimental ΔΔG values of 0.58 in cross validation and of 0.72 on

a blind test set. The sequence-based INPS scores slightly lower than the

structure-based INPS3D and both on the same blind test sets well compare with the

state-of-the-art methods.

AVAILABILITY AND IMPLEMENTATION: INPS and INPS3D are available at the same web

server: http://inpsmd.biocomp.unibo.it

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

CONTACT: gigi@biocomp.unibo.it.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw192

PMID: 27153629

130. Bioinformatics. 2016 Aug 15;32(16):2411-8. doi: 10.1093/bioinformatics/btw186.

Epub 2016 Apr 8.

iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes

of pseudo nucleotide composition into an ensemble learning framework.

Liu B(1), Long R(2), Chou KC(3).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong 518055, China Key Laboratory of

Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, Guangdong 518055, China Gordon Life Science Institute,

Belmont, MA 02478, USA. (2)School of Computer Science and Technology, Harbin

Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055,

China. (3)Gordon Life Science Institute, Belmont, MA 02478, USA Center of

Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University,

Jeddah 21589, Saudi Arabia.

MOTIVATION: Regulatory DNA elements are associated with DNase I hypersensitive

sites (DHSs). Accordingly, identification of DHSs will provide useful insights

for in-depth investigation into the function of noncoding genomic regions.

RESULTS: In this study, using the strategy of ensemble learning framework, we

proposed a new predictor called iDHS-EL for identifying the location of DHS in

human genome. It was formed by fusing three individual Random Forest (RF)

classifiers into an ensemble predictor. The three RF operators were respectively

based on the three special modes of the general pseudo nucleotide composition

(PseKNC): (i) kmer, (ii) reverse complement kmer and (iii) pseudo dinucleotide

composition. It has been demonstrated that the new predictor remarkably

outperforms the relevant state-of-the-art methods in both accuracy and stability.

AVAILABILITY AND IMPLEMENTATION: For the convenience of most experimental

scientists, a web server for iDHS-EL is established at

http://bioinformatics.hitsz.edu.cn/iDHS-EL, which is the first web-server

predictor ever established for identifying DHSs, and by which users can easily

get their desired results without the need to go through the mathematical

details. We anticipate that IDHS-EL: will become a very useful high throughput

tool for genome analysis.

CONTACT: bliu@gordonlifescience.org or bliu@insun.hit.edu.cn

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw186

PMID: 27153623

131. Bioinformatics. 2016 Aug 15;32(16):2502-4. doi: 10.1093/bioinformatics/btw157.

Epub 2016 Apr 7.

MGEScan: a Galaxy-based system for identifying retrotransposons in genomes.

Lee H(1), Lee M(2), Mohammed Ismail W(1), Rho M(3), Fox GC(1), Oh S(4), Tang

H(1).

Author information:

(1)School of Informatics and Computing, Indiana University, Bloomington, IN, USA.

(2)Department of Computer Science and Engineering, Ewha Womans University, Seoul,

Korea. (3)Department of Computer Science and Engineering, Hanyang University,

Seoul, Korea. (4)Department of Software Convergence Technology, Ajou University,

Suwon, Korea.

: MGEScan-long terminal repeat (LTR) and MGEScan-non-LTR are successfully used

programs for identifying LTRs and non-LTR retrotransposons in eukaryotic genome

sequences. However, these programs are not supported by easy-to-use interfaces

nor well suited for data visualization in general data formats. Here, we present

MGEScan, a user-friendly system that combines these two programs with a Galaxy

workflow system accelerated with MPI and Python threading on compute clusters.

MGEScan and Galaxy empower researchers to identify transposable elements in a

graphical user interface with ready-to-use workflows. MGEScan also visualizes the

custom annotation tracks for mobile genetic elements in public genome browsers. A

maximum speed-up of 3.26× is attained for execution time using concurrent

processing and MPI on four virtual cores. MGEScan provides four operational

modes: as a command line tool, as a Galaxy Toolshed, on a Galaxy-based web

server, and on a virtual cluster on the Amazon cloud.AVAILABILITY AND

IMPLEMENTATION: MGEScan tutorials and source code are available at

http://mgescan.readthedocs.org/

CONTACT: hatang@indiana.edu or syoh@ajou.ac.kr

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw157

PMID: 27153595

132. PLoS One. 2016 Aug 15;11(8):e0155290. doi: 10.1371/journal.pone.0155290.

eCollection 2016.

SVM-Prot 2016: A Web-Server for Machine Learning Prediction of Protein Functional

Families from Sequence Irrespective of Similarity.

Li YH(1), Xu JY(1,)(2), Tao L(1,)(3), Li XF(1), Li S(1), Zeng X(3), Chen SY(3),

Zhang P(3), Qin C(3), Zhang C(3), Chen Z(4), Zhu F(1), Chen YZ(3).

Author information:

(1)Innovative Drug Research and Bioinformatics Group, Innovative Drug Research

Centre and School of Pharmaceutical Sciences, Chongqing University, Chongqing,

401331, China. (2)School of Mathematics and Statistics, Beijing Institute of

Technology, Beijing, China. (3)Bioinformatics and Drug Discovery group,

Department of Pharmacy, National University of Singapore, Singapore, 117543,

Singapore. (4)Zhejiang Key Laboratory of Gastro-intestinal Pathophysiology,

Zhejiang Hospital of Traditional Chinese Medicine, Zhejiang Chinese Medical

University, Hangzhou, P. R. China.

Knowledge of protein function is important for biological, medical and

therapeutic studies, but many proteins are still unknown in function. There is a

need for more improved functional prediction methods. Our SVM-Prot web-server

employed a machine learning method for predicting protein functional families

from protein sequences irrespective of similarity, which complemented those

similarity-based and other methods in predicting diverse classes of proteins

including the distantly-related proteins and homologous proteins of different

functions. Since its publication in 2003, we made major improvements to SVM-Prot

with (1) expanded coverage from 54 to 192 functional families, (2) more diverse

protein descriptors protein representation, (3) improved predictive performances

due to the use of more enriched training datasets and more variety of protein

descriptors, (4) newly integrated BLAST analysis option for assessing proteins in

the SVM-Prot predicted functional families that were similar in sequence to a

query protein, and (5) newly added batch submission option for supporting the

classification of multiple proteins. Moreover, 2 more machine learning

approaches, K nearest neighbor and probabilistic neural networks, were added for

facilitating collective assessment of protein functions by multiple methods.

SVM-Prot can be accessed at http://bidd2.nus.edu.sg/cgi-bin/svmprot/svmprot.cgi.

DOI: 10.1371/journal.pone.0155290

PMCID: PMC4985167

PMID: 27525735

133. Biochem Biophys Res Commun. 2016 Aug 12;477(1):150-4. doi:

10.1016/j.bbrc.2016.06.035. Epub 2016 Jun 10.

Prediction of cell-penetrating peptides with feature selection techniques.

Tang H(1), Su ZD(2), Wei HH(2), Chen W(3), Lin H(4).

Author information:

(1)Department of Pathophysiology, Southwest Medical University, Luzhou 646000,

China. Electronic address: tanghua771211@aliyun.com. (2)Key Laboratory for

NeuroInformation of Ministry of Education, Center of Bioinformatics and Center

for Information in Biomedicine, School of Life Science and Technology, University

of Electronic Science and Technology of China, Chengdu 610054, China.

(3)Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, North China University of Science and Technology, Tangshan

063009, China; Key Laboratory for NeuroInformation of Ministry of Education,

Center of Bioinformatics and Center for Information in Biomedicine, School of

Life Science and Technology, University of Electronic Science and Technology of

China, Chengdu 610054, China. (4)Key Laboratory for NeuroInformation of Ministry

of Education, Center of Bioinformatics and Center for Information in Biomedicine,

School of Life Science and Technology, University of Electronic Science and

Technology of China, Chengdu 610054, China. Electronic address:

hlin@uestc.edu.cn.

Cell-penetrating peptides are a group of peptides which can transport different

types of cargo molecules such as drugs across plasma membrane and have been

applied in the treatment of various diseases. Thus, the accurate prediction of

cell-penetrating peptides with bioinformatics methods will accelerate the

development of drug delivery systems. The study aims to develop a powerful model

to accurately identify cell-penetrating peptides. At first, the peptides were

translated into a set of vectors with the same dimension by using dipeptide

compositions. Secondly, the Analysis of Variance-based technique was used to

reduce the dimension of the vector and explore the optimized features. Finally,

the support vector machine was utilized to discriminate cell-penetrating peptides

from non-cell-penetrating peptides. The five-fold cross-validated results showed

that our proposed method could achieve an overall prediction accuracy of 83.6%.

Based on the proposed model, we constructed a free webserver called C2Pred

(http://lin.uestc.edu.cn/server/C2Pred).

Copyright © 2016 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.bbrc.2016.06.035

PMID: 27291150

134. BMC Genomics. 2016 Aug 12;17(1):632. doi: 10.1186/s12864-016-2963-0.

iTAR: a web server for identifying target genes of transcription factors using

ChIP-seq or ChIP-chip data.

Yang CC(1,)(2), Andrews EH(3), Chen MH(2,)(4), Wang WY(2), Chen JJ(1,)(4,)(5),

Gerstein M(6,)(7), Liu CC(8,)(9,)(10), Cheng C(11,)(12,)(13).

Author information:

(1)Institute of Molecular Biology, National Chung Hsing University, Taichung,

402, Taiwan. (2)Institute of Genomics and Bioinformatics, National Chung Hsing

University, Taichung, 402, Taiwan. (3)Geisel School of Medicine at Dartmouth,

Institute for Quantitative Biomedical Sciences, Lebanon, NH, 03766, USA.

(4)Institute of Biomedical Sciences, National Chung Hsing University, Taichung,

402, Taiwan. (5)Agricultural Biotechnology Center, National Chung Hsing

University, Taichung, 402, Taiwan. (6)Program in Computational Biology and

Bioinformatics, Yale University, 260 Whitney Avenue, New Haven, CT, 06520, USA.

(7)Department of Molecular Biophysics and Biochemistry, Yale University, 260

Whitney Avenue, New Haven, CT, 06520, USA. (8)Institute of Genomics and

Bioinformatics, National Chung Hsing University, Taichung, 402, Taiwan.

jimliu@nchu.edu.tw. (9)Institute of Biomedical Sciences, National Chung Hsing

University, Taichung, 402, Taiwan. jimliu@nchu.edu.tw. (10)Agricultural

Biotechnology Center, National Chung Hsing University, Taichung, 402, Taiwan.

jimliu@nchu.edu.tw. (11)Department of Genetics, Geisel School of Medicine at

Dartmouth, Hanover, NH, 03755, USA. chao.cheng@dartmouth.edu. (12)Geisel School

of Medicine at Dartmouth, Institute for Quantitative Biomedical Sciences,

Lebanon, NH, 03766, USA. chao.cheng@dartmouth.edu. (13)Norris Cotton Cancer

Center, Geisel School of Medicine at Dartmouth, Lebanon, NH, 03766, USA.

chao.cheng@dartmouth.edu.

BACKGROUND: Chromatin immunoprecipitation followed by massively parallel DNA

sequencing (ChIP-seq) or microarray hybridization (ChIP-chip) has been widely

used to determine the genomic occupation of transcription factors (TFs). We have

previously developed a probabilistic method, called TIP (Target Identification

from Profiles), to identify TF target genes using ChIP-seq/ChIP-chip data. To

achieve high specificity, TIP applies a conservative method to estimate

significance of target genes, with the trade-off being a relatively low

sensitivity of target gene identification compared to other methods.

Additionally, TIP's output does not render binding-peak locations or intensity,

information highly useful for visualization and general experimental biological

use, while the variability of ChIP-seq/ChIP-chip file formats has made input into

TIP more difficult than desired.

DESCRIPTION: To improve upon these facets, here we present are fined TIP with key

extensions. First, it implements a Gaussian mixture model for p-value estimation,

increasing target gene identification sensitivity and more accurately capturing

the shape of TF binding profile distributions. Second, it enables the

incorporation of TF binding-peak data by identifying their locations in

significant target gene promoter regions and quantifies their strengths. Finally,

for full ease of implementation we have incorporated it into a web server (

http://syslab3.nchu.edu.tw/iTAR/ ) that enables flexibility of input file format,

can be used across multiple species and genome assembly versions, and is freely

available for public use. The web server additionally performs GO enrichment

analysis for the identified target genes to reveal the potential function of the

corresponding TF.

CONCLUSIONS: The iTAR web server provides a user-friendly interface and supports

target gene identification in seven species, ranging from yeast to human. To

facilitate investigating the quality of ChIP-seq/ChIP-chip data, the web server

generates the chart of the characteristic binding profiles and the density plot

of normalized regulatory scores. The iTAR web server is a useful tool in

identifying TF target genes from ChIP-seq/ChIP-chip data and discovering

biological insights.

DOI: 10.1186/s12864-016-2963-0

PMCID: PMC4983039

PMID: 27519564

135. BMC Genomics. 2016 Aug 11;17(1):622. doi: 10.1186/s12864-016-2964-z.

COMAN: a web server for comprehensive metatranscriptomics analysis.

Ni Y(1), Li J(1), Panagiotou G(2).

Author information:

(1)Systems Biology & Bioinformatics Group, School of Biological Sciences, The

University of Hong Kong, Pokfulam Road, Hong Kong, Hong Kong. (2)Systems Biology

& Bioinformatics Group, School of Biological Sciences, The University of Hong

Kong, Pokfulam Road, Hong Kong, Hong Kong. gipa@hku.hk.

BACKGROUND: Microbiota-oriented studies based on metagenomic or

metatranscriptomic sequencing have revolutionised our understanding on microbial

ecology and the roles of both clinical and environmental microbes. The analysis

of massive metatranscriptomic data requires extensive computational resources, a

collection of bioinformatics tools and expertise in programming.

RESULTS: We developed COMAN (Comprehensive Metatranscriptomics Analysis), a

web-based tool dedicated to automatically and comprehensively analysing

metatranscriptomic data. COMAN pipeline includes quality control of raw reads,

removal of reads derived from non-coding RNA, followed by functional annotation,

comparative statistical analysis, pathway enrichment analysis, co-expression

network analysis and high-quality visualisation. The essential data generated by

COMAN are also provided in tabular format for additional analysis and integration

with other software. The web server has an easy-to-use interface and detailed

instructions, and is freely available at http://sbb.hku.hk/COMAN/ CONCLUSIONS:

COMAN is an integrated web server dedicated to comprehensive functional analysis

of metatranscriptomic data, translating massive amount of reads to data tables

and high-standard figures. It is expected to facilitate the researchers with less

expertise in bioinformatics in answering microbiota-related biological questions

and to increase the accessibility and interpretation of microbiota RNA-Seq data.

DOI: 10.1186/s12864-016-2964-z

PMCID: PMC4982211

PMID: 27515514

136. Sci Rep. 2016 Aug 11;6:31080. doi: 10.1038/srep31080.

RAMPred: identifying the N(1)-methyladenosine sites in eukaryotic transcriptomes.

Chen W(1), Feng P(2), Tang H(3), Ding H(4), Lin H(4).

Author information:

(1)Department of Physics, School of Sciences, and Center for Genomics and

Computational Biology, North China University of Science and Technology,

Tangshan, Tangshan 063000, China. (2)School of Public Health, North China

University of Science and Technology, Tangshan, 063000, China. (3)Department of

Pathophysiology, Southwest Medical University, Luzhou 646000, China. (4)Key

Laboratory for Neuro-Information of Ministry of Education, Center of

Bioinformatics and Center for Information in Biomedicine, School of Life Science

and Technology, University of Electronic Science and Technology of China, Chengdu

610054, China.

N(1)-methyladenosine (m(1)A) is a prominent RNA modification involved in many

biological processes. Accurate identification of m(1)A site is invaluable for

better understanding the biological functions of m(1)A. However, limitations in

experimental methods preclude the progress towards the identification of m(1)A

site. As an excellent complement of experimental methods, a support vector

machine based-method called RAMPred is proposed to identify m(1)A sites in H.

sapiens, M. musculus and S. cerevisiae genomes for the first time. In this

method, RNA sequences are encoded by using nucleotide chemical property and

nucleotide compositions. RAMPred achieves promising performances in jackknife

tests, cross cell line tests and cross species tests, indicating that RAMPred

holds very high potential to become a useful tool for identifying m(1)A sites.

For the convenience of experimental scientists, a web-server based on the

proposed model was constructed and could be freely accessible at

http://lin.uestc.edu.cn/server/RAMPred.

DOI: 10.1038/srep31080

PMCID: PMC4980636

PMID: 27511610

137. Oncotarget. 2016 Aug 9;7(32):51270-51283. doi: 10.18632/oncotarget.9987.

iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different

pseudo components into an ensemble classifier.

Qiu WR(1,)(2), Xiao X(1,)(3), Xu ZC(1), Chou KC(3,)(4,)(5).

Author information:

(1)Computer Department, Jingdezhen Ceramic Institute, Jingdezhen, China.

(2)Department of Computer Science and Bond Life Science Center, University of

Missouri, Columbia, MO, USA. (3)Gordon Life Science Institute, Boston, MA, USA.

(4)Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz

University, Jeddah, Saudi Arabia. (5)Center of Bioinformatics, School of Life

Science and Technology, University of Electronic Science and Technology of China,

Chengdu, Sichuan, China.

Protein phosphorylation is a posttranslational modification (PTM or PTLM), where

a phosphoryl group is added to the residue(s) of a protein molecule. The most

commonly phosphorylated amino acids occur at serine (S), threonine (T), and

tyrosine (Y). Protein phosphorylation plays a significant role in a wide range of

cellular processes; meanwhile its dysregulation is also involved with many

diseases. Therefore, from the angles of both basic research and drug development,

we are facing a challenging problem: for an uncharacterized protein sequence

containing many residues of S, T, or Y, which ones can be phosphorylated, and

which ones cannot? To address this problem, we have developed a predictor called

iPhos-PseEn by fusing four different pseudo component approaches (amino acids'

disorder scores, nearest neighbor scores, occurrence frequencies, and position

weights) into an ensemble classifier via a voting system. Rigorous

cross-validations indicated that the proposed predictor remarkably outperformed

its existing counterparts. For the convenience of most experimental scientists, a

user-friendly web-server for iPhos-PseEn has been established at

http://www.jci-bioinfo.cn/iPhos-PseEn, by which users can easily obtain their

desired results without the need to go through the complicated mathematical

equations involved.

DOI: 10.18632/oncotarget.9987

PMCID: PMC5239474

PMID: 27323404

Conflict of interest statement: The authors declare no conflicts of interest.

138. Bioinformatics. 2016 Aug 1;32(15):2256-63. doi: 10.1093/bioinformatics/btw169.

Epub 2016 Apr 1.

GenomeRunner web server: regulatory similarity and differences define the

functional impact of SNP sets.

Dozmorov MG(1), Cara LR(1), Giles CB(2), Wren JD(3).

Author information:

(1)Department of Biostatistics, Virginia Commonwealth University, Richmond, VA,

USA. (2)Department of Arthritis and Clinical Immunology, Oklahoma Medical

Research Foundation, Oklahoma City, OK, USA. (3)Department of Arthritis and

Clinical Immunology, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA

Department of Biochemistry and Molecular Biology, University of Oklahoma Health

Sciences Center, Oklahoma City, OK, USA.

MOTIVATION: The growing amount of regulatory data from the ENCODE, Roadmap

Epigenomics and other consortia provides a wealth of opportunities to investigate

the functional impact of single nucleotide polymorphisms (SNPs). Yet, given the

large number of regulatory datasets, researchers are posed with a challenge of

how to efficiently utilize them to interpret the functional impact of SNP sets.

RESULTS: We developed the GenomeRunner web server to automate systematic

statistical analysis of SNP sets within a regulatory context. Besides defining

the functional impact of SNP sets, GenomeRunner implements novel regulatory

similarity/differential analyses, and cell type-specific regulatory enrichment

analysis. Validated against literature- and disease ontology-based approaches,

analysis of 39 disease/trait-associated SNP sets demonstrated that the functional

impact of SNP sets corresponds to known disease relationships. We identified a

group of autoimmune diseases with SNPs distinctly enriched in the enhancers of T

helper cell subpopulations, and demonstrated relevant cell type-specificity of

the functional impact of other SNP sets. In summary, we show how systematic

analysis of genomic data within a regulatory context can help interpreting the

functional impact of SNP sets.

AVAILABILITY AND IMPLEMENTATION: GenomeRunner web server is freely available at

http://www.integrativegenomics.org/

CONTACT: mikhail.dozmorov@gmail.com

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw169

PMCID: PMC4965636 [Available on 2017-08-01]

PMID: 27153607

139. Bioinformatics. 2016 Aug 1;32(15):2386-8. doi: 10.1093/bioinformatics/btw141.

Epub 2016 Mar 12.

FRODOCK 2.0: fast protein-protein docking server.

Ramírez-Aportela E(1), López-Blanco JR(2), Chacón P(2).

Author information:

(1)Department of Biological Chemical Physics, Rocasolano Physical Chemistry

Institute C.S.I.C, Madrid 28006, Spain Centro de Investigaciones Biológicas,

CSIC, Madrid E-28040, Spain. (2)Department of Biological Chemical Physics,

Rocasolano Physical Chemistry Institute C.S.I.C, Madrid 28006, Spain.

The prediction of protein-protein complexes from the structures of unbound

components is a challenging and powerful strategy to decipher the mechanism of

many essential biological processes. We present a user-friendly protein-protein

docking server based on an improved version of FRODOCK that includes a

complementary knowledge-based potential. The web interface provides a very

effective tool to explore and select protein-protein models and interactively

screen them against experimental distance constraints. The competitive success

rates and efficiency achieved allow the retrieval of reliable potential

protein-protein binding conformations that can be further refined with more

computationally demanding strategies.AVAILABILITY AND IMPLEMENTATION: The server

is free and open to all users with no login requirement at

http://frodock.chaconlab.org

CONTACT: pablo@chaconlab.org

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw141

PMID: 27153583

140. Bioinformatics. 2016 Aug 1;32(15):2289-96. doi: 10.1093/bioinformatics/btw133.

Epub 2016 Mar 11.

PinaColada: peptide-inhibitor ant colony ad-hoc design algorithm.

Zaidman D(1), Wolfson HJ(1).

Author information:

(1)Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978,

Israel.

MOTIVATION: Design of protein-protein interaction (PPI) inhibitors is a major

challenge in Structural Bioinformatics. Peptides, especially short ones (5-15

amino acid long), are natural candidates for inhibition of protein-protein

complexes due to several attractive features such as high structural

compatibility with the protein binding site (mimicking the surface of one of the

proteins), small size and the ability to form strong hotspot binding connections

with the protein surface. Efficient rational peptide design is still a major

challenge in computer aided drug design, due to the huge space of possible

sequences, which is exponential in the length of the peptide, and the high

flexibility of peptide conformations.

RESULTS: In this article we present PinaColada, a novel computational method for

the design of peptide inhibitors for protein-protein interactions. We employ a

version of the ant colony optimization heuristic, which is used to explore the

exponential space ([Formula: see text]) of length n peptide sequences, in

combination with our fast robotics motivated PepCrawler algorithm, which explores

the conformational space for each candidate sequence. PinaColada is being run in

parallel, on a DELL PowerEdge 2.8 GHZ computer with 20 cores and 256 GB memory,

and takes up to 24 h to design a peptide of 5-15 amino acids length.

AVAILABILITY AND IMPLEMENTATION: An online server available at:

http://bioinfo3d.cs.tau.ac.il/PinaColada/.

CONTACT: danielza@post.tau.ac.il; wolfson@tau.ac.il.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw133

PMID: 27153578

141. Curr Protoc Immunol. 2016 Aug 1;114:18.19.1-18.19.24. doi: 10.1002/cpim.12.

TepiTool: A Pipeline for Computational Prediction of T Cell Epitope Candidates.

Paul S(1), Sidney J(1), Sette A(1), Peters B(1).

Author information:

(1)La Jolla Institute for Allergy and Immunology, Division of Vaccine Discovery,

La Jolla, California.

Computational prediction of T cell epitope candidates is currently being used in

several applications including vaccine discovery studies, development of

diagnostics, and removal of unwanted immune responses against protein

therapeutics. There have been continuous improvements in the performance of MHC

binding prediction tools, but their general adoption by immunologists has been

slow due to the lack of user-friendly interfaces and guidelines. Current tools

only provide minimal advice on what alleles to include, what lengths to consider,

how to deal with homologous peptides, and what cutoffs should be considered

relevant. This protocol provides step-by-step instructions with necessary

recommendations for prediction of the best T cell epitope candidates with the

newly developed online tool called TepiTool. TepiTool, which is part of the

Immune Epitope Database (IEDB), provides some of the top MHC binding prediction

algorithms for number of species including humans, chimpanzees, bovines,

gorillas, macaques, mice, and pigs. The TepiTool is freely accessible at

http://tools.iedb.org/tepitool/. © 2016 by John Wiley & Sons, Inc.

Copyright © 2016 John Wiley & Sons, Inc.

DOI: 10.1002/cpim.12

PMCID: PMC4981331 [Available on 2017-08-01]

PMID: 27479659

142. Hum Mutat. 2016 Aug;37(8):820-6. doi: 10.1002/humu.23007. Epub 2016 May 24.

Enrichment of SNPs in Functional Categories Reveals Genes Affecting Complex

Traits.

Zhao H(1), Fan D(2), Nyholt DR(1), Yang Y(3).

Author information:

(1)Institute of Health and Biomedical Innovation, Queensland University of

Technology, Queensland, Australia. (2)Department of Neurology, Peking University

Third Hospital, Beijing, China. (3)Institute for Glycomics, Griffith University,

Queensland, Australia.

Genome-wide association studies (GWAS) have indicated potential to identify

heritability of common complex phenotypes, but traditional approaches have

limited ability to detect hiding signals because single SNP has weak effect size

accounting for only a small fraction of overall phenotypic variations. To improve

the power of GWAS, methods have been developed to identify truly associated genes

by jointly testing effects of all SNPs. However, equally considering all SNPs

within a gene might dilute strong signals of SNPs in real functional categories.

Here, we observed a consistent pattern on enrichment of significant SNPs in eight

functional categories across six phenotypes, with the highest enrichment in

coding and both UTR regions while the lowest enrichment in the intron. Based on

the pattern of SNP enrichment in functional categories, we developed a new

approach for detecting gene associations on traits (DGAT) by selecting the most

significant functional category and then using SNPs within it to assess gene

associations. The method was found to be robust in type I error rate on simulated

data, and to have mostly higher power in detecting associated genes for three

different diseases than other methods. Further analysis indicated ability of the

DGAT to detect novel genes. The DGAT is available by

http://sparks-lab.org/server/DGAT.

© 2016 WILEY PERIODICALS, INC.

DOI: 10.1002/humu.23007

PMID: 27113629

143. Neurol Clin Pract. 2016 Aug;6(4):304-314.

Knowledge translation of an online tool to determine candidacy for epilepsy

surgery evaluation.

Sauro KM(1), Holroyd-Leduc J(1), Wiebe S(1), Quan H(1), Cooke L(1), Cross JH(1),

Mathern GW(1), Armson H(1), Stromer J(1), Jetté N(1).

Author information:

(1)Department of Clinical Neurosciences & Hotchkiss Brain Institute (KMS, SW, LC,

NJ), Department of Community Health Sciences & O'Brien Institute for Public

Health (KMS, JH-L, SW, HQ, NJ), Department of Medicine (JH-L), Department of

Family Medicine (HA), and Ward of the 21st Century (JS), Cumming School of

Medicine, University of Calgary, Alberta, Canada; UCL Institute of Child Health &

Great Ormond Street Hospital for Children (JHC), London, UK; and Department of

Neurosurgery, Ronald Reagan UCLA Medical Center, University of California (GWM),

Los Angeles.

BACKGROUND: Guidelines recommend that surgery be considered in patients with

drug-resistant epilepsy, yet delays to epilepsy surgery still exist. A Web-based,

evidence-informed clinical decision tool (www.toolsforepilepsy.com) was developed

to help physicians determine which patients are appropriate for an epilepsy

surgery evaluation. We evaluated the usability and feasibility of the tool with

the intended end users in order to improve implementation into practice.

METHODS: Usability testing was conducted with relevant end users. After the tool

was modified based on usability results, another group of end users trialed the

tool in their clinical practice. This latter group of end users then participated

in focus groups and semi-structured interviews to address barriers and

facilitators to tool implementation. Finally, a stakeholder meeting was held with

domain experts and end users to discuss further changes to the tool and

implementation strategies.

RESULTS: Six overall themes were identified through usability testing, and an

additional 11 themes were identified through the focus groups and interviews. The

tool was modified based on these findings, which were then presented at the

stakeholder meeting of experts and end users for further refinement. The findings

were also used to guide discussions of potential implementation strategies at the

meeting.

CONCLUSION: This study provides guidance on how to improve the usability of

clinical decision tools by engaging end users, experts, and other key

stakeholders. The modifications to the tool should facilitate its implementation

in clinical practice and ultimately enhance the quality of care persons with

epilepsy receive.

DOI: 10.1212/CPJ.0000000000000250

PMCID: PMC4987121 [Available on 2017-08-01]

PMID: 27574569

144. Protein Eng Des Sel. 2016 Aug;29(8):281-4. doi: 10.1093/protein/gzw021. Epub 2016

Jun 9.

ProSAT+: visualizing sequence annotations on 3D structure.

Stank A(1), Richter S(2), Wade RC(3).

Author information:

(1)Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg

35, Heidelberg 69118, Germany Heidelberg Graduate School of Mathematical and

Computational Methods for the Sciences, Im Neuenheimer Feld 205, Heidelberg

69120, Germany. (2)Heidelberg Institute for Theoretical Studies (HITS),

Schloss-Wolfsbrunnenweg 35, Heidelberg 69118, Germany. (3)Heidelberg Institute

for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, Heidelberg 69118,

Germany Center for Molecular Biology at Heidelberg University (ZMBH), DKFZ-ZMBH

Alliance and Interdisciplinary Center for Scientific Computing (IWR), Im

Neuenheimer Feld 282, Heidelberg 69120, Germany rebecca.wade@h-its.org.

PRO: tein S: tructure A: nnotation T: ool-plus (ProSAT(+)) is a new web server

for mapping protein sequence annotations onto a protein structure and visualizing

them simultaneously with the structure. ProSAT(+) incorporates many of the

features of the preceding ProSAT and ProSAT2 tools but also provides new options

for the visualization and sharing of protein annotations. Data are extracted from

the UniProt KnowledgeBase, the RCSB PDB and the PDBe SIFTS resource, and

visualization is performed using JSmol. User-defined sequence annotations can be

added directly to the URL, thus enabling visualization and easy data sharing.

ProSAT(+) is available at http://prosat.h-its.org.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/protein/gzw021

PMID: 27284084

145. PLoS Comput Biol. 2016 Jul 29;12(7):e1004976. doi: 10.1371/journal.pcbi.1004976.

eCollection 2016.

PhyloBot: A Web Portal for Automated Phylogenetics, Ancestral Sequence

Reconstruction, and Exploration of Mutational Trajectories.

Hanson-Smith V(1), Johnson A(1).

Author information:

(1)Department of Microbiology and Immunology, University of California, San

Francisco, San Francisco, California, United States of America.

The method of phylogenetic ancestral sequence reconstruction is a powerful

approach for studying evolutionary relationships among protein sequence,

structure, and function. In particular, this approach allows investigators to (1)

reconstruct and "resurrect" (that is, synthesize in vivo or in vitro) extinct

proteins to study how they differ from modern proteins, (2) identify key amino

acid changes that, over evolutionary timescales, have altered the function of the

protein, and (3) order historical events in the evolution of protein function.

Widespread use of this approach has been slow among molecular biologists, in part

because the methods require significant computational expertise. Here we present

PhyloBot, a web-based software tool that makes ancestral sequence reconstruction

easy. Designed for non-experts, it integrates all the necessary software into a

single user interface. Additionally, PhyloBot provides interactive tools to

explore evolutionary trajectories between ancestors, enabling the rapid

generation of hypotheses that can be tested using genetic or biochemical

approaches. Early versions of this software were used in previous studies to

discover genetic mechanisms underlying the functions of diverse protein families,

including V-ATPase ion pumps, DNA-binding transcription regulators, and

serine/threonine protein kinases. PhyloBot runs in a web browser, and is

available at the following URL: http://www.phylobot.com. The software is

implemented in Python using the Django web framework, and runs on elastic cloud

computing resources from Amazon Web Services. Users can create and submit jobs on

our free server (at the URL listed above), or use our open-source code to launch

their own PhyloBot server.

DOI: 10.1371/journal.pcbi.1004976

PMCID: PMC4966924

PMID: 27472806

146. Oncotarget. 2016 Jul 28. doi: 10.18632/oncotarget.10886. [Epub ahead of print]

DNA-damage related genes and clinical outcome in hormone receptor positive breast

cancer.

Nieto-Jiménez C(1), Alcaraz-Sanabria A(1), Páez R(1), Pérez-Peña J(1),

Corrales-Sánchez V(1), Pandiella A(2), Ocaña A(1).

Author information:

(1)Translational Research Unit, Albacete University Hospital, Albacete, Spain.

(2)Cancer Research Center, CSIC-University of Salamanca, Salamanca, Spain.

BACKGROUND: Control of DNA damage is frequently deregulated in solid tumors.

Upregulation of genes within this process can be indicative of a more aggressive

phenotype and linked with worse outcome. In the present article we identify DNA

damage related genes associated with worse outcome in breast cancer.

RESULTS: 2286 genes were differentially expressed between normal breast tissue

and basal-like tumors, and 62 included in the DNA metabolic process function.

Expression of RAD51, GINS1, TRIP13 and MCM2 were associated with detrimental

relapse free survival (RFS) and overall survival (OS) in luminal tumors. The

combined analyses of TRIP13+RAD51+MCM2 showed the worse association for RFS (HR

2.25 (1.51-3.35) log rank p= 4.1e-05) and TRIP13+RAD51 for OS (HR 5.13

(0.6-44.17) log rank p=0.098) in ER+/HER2- tumors. TRIP13 is amplified in 3.1% of

breast cancers.

METHODS: Transcriptomic analyses using public datasets evaluating expression

values between normal breast tissue and TNBC identified upregulated genes. Genes

included in the DNA metabolic process were selected and confirmed using data

contained at oncomine (www.oncomine.org). Evaluation of the selected genes with

RFS and OS was performed using the KM Plotter Online Tool

(http://www.kmplot.com). Evaluation of molecular alterations was performed using

cBioportal (www.cbioportal.org).

CONCLUSIONS: Expression of DNA metabolic related genes RAD51, GINS1, TRIP13 and

MCM2 are associated with poor outcome. Combinations of some of these genes are

linked to poor RFS or OS in luminal A, B and ER+HER2- tumors. Evaluation of its

predictive capacity in prospective studies is required.

DOI: 10.18632/oncotarget.10886

PMID: 27486818

147. Bioinformatics. 2016 Jul 15;32(14):2224-6. doi: 10.1093/bioinformatics/btw147.

Epub 2016 Mar 18.

Ferret: a user-friendly Java tool to extract data from the 1000 Genomes Project.

Limou S(1), Taverner AM(2), Winkler CA(1).

Author information:

(1)Molecular Genetic Epidemiology Section, Basic Research Laboratory, Basic

Science Program, NCI, Leidos Biomedical Research, Inc., Frederick National

Laboratory, Frederick, MD 21702, USA. (2)Molecular Genetic Epidemiology Section,

Basic Research Laboratory, Basic Science Program, NCI, Leidos Biomedical

Research, Inc., Frederick National Laboratory, Frederick, MD 21702, USA

Quantitative and Computational Biology program, Princeton University, Princeton,

NJ 08544, USA.

The 1000 Genomes (1KG) Project provides a near-comprehensive resource on human

genetic variation in worldwide reference populations. 1KG variants can be

accessed through a browser and through the raw and annotated data that are

regularly released on an ftp server. We developed Ferret, a user-friendly Java

tool, to easily extract genetic variation information from these large and

complex data files. From a locus, gene(s) or SNP(s) of interest, Ferret retrieves

genotype data for 1KG SNPs and indels, and computes allelic frequencies for 1KG

populations and optionally, for the Exome Sequencing Project populations. By

converting the 1KG data into files that can be imported into popular pre-existing

tools (e.g. PLINK and HaploView), Ferret offers a straightforward way, even for

non-bioinformatics specialists, to manipulate, explore and merge 1KG data with

the user's dataset, as well as visualize linkage disequilibrium pattern, infer

haplotypes and design tagSNPs.AVAILABILITY AND IMPLEMENTATION: Ferret tool and

source code are publicly available at http://limousophie35.github.io/Ferret/

CONTACT: ferret@nih.gov

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

Published by Oxford University Press 2016. This work is written by US Government

employees and is in the public domain in the US.

DOI: 10.1093/bioinformatics/btw147

PMCID: PMC4937191 [Available on 2017-07-15]

PMID: 27153588

148. Oncotarget. 2016 Jul 12;7(28):44310-44321. doi: 10.18632/oncotarget.10027.

iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by

incorporating sequence-coupled effects into general PseAAC.

Qiu WR(1,)(2), Sun BQ(1), Xiao X(1,)(3), Xu ZC(1), Chou KC(3,)(4,)(5).

Author information:

(1)Computer Department, Jingdezhen Ceramic Institute, Jingdezhen, China.

(2)Department of Computer Science and Bond Life Science Center, University of

Missouri, Columbia, MO, USA. (3)Gordon Life Science Institute, Boston, MA, USA.

(4)Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz

University, Jeddah, Saudi Arabia. (5)Center of Bioinformatics, School of Life

Science and Technology, University of Electronic Science and Technology of China,

Chengdu, Sichuan, China.

Protein hydroxylation is a posttranslational modification (PTM), in which a CH

group in Pro (P) or Lys (K) residue has been converted into a COH group, or a

hydroxyl group (-OH) is converted into an organic compound. Closely associated

with cellular signaling activities, this type of PTM is also involved in some

major diseases, such as stomach cancer and lung cancer. Therefore, from the

angles of both basic research and drug development, we are facing a challenging

problem: for an uncharacterized protein sequence containing many residues of P or

K, which ones can be hydroxylated, and which ones cannot? With the explosive

growth of protein sequences in the post-genomic age, the problem has become even

more urgent. To address such a problem, we have developed a predictor called

iHyd-PseCp by incorporating the sequence-coupled information into the general

pseudo amino acid composition (PseAAC) and introducing the "Random Forest"

algorithm to operate the calculation. Rigorous jackknife tests indicated that the

new predictor remarkably outperformed the existing state-of-the-art prediction

method for the same purpose. For the convenience of most experimental scientists,

a user-friendly web-server for iHyd-PseCp has been established at

http://www.jci-bioinfo.cn/iHyd-PseCp, by which users can easily obtain their

desired results without the need to go through the complicated mathematical

equations involved.

DOI: 10.18632/oncotarget.10027

PMCID: PMC5190098

PMID: 27322424

Conflict of interest statement: The authors declare no conflicts of interest.

149. J Cheminform. 2016 Jul 11;8:38. doi: 10.1186/s13321-016-0149-z. eCollection 2016.

bSiteFinder, an improved protein-binding sites prediction server based on

structural alignment: more accurate and less time-consuming.

Gao J(1), Zhang Q(2), Liu M(3), Zhu L(4), Wu D(2), Cao Z(2), Zhu R(2).

Author information:

(1)Department of Bioinformatics, Tongji University, Shanghai, 200092 People's

Republic of China ; School of Information Engineering, Shanghai Maritime

University, Shanghai, 201306 People's Republic of China. (2)Department of

Bioinformatics, Tongji University, Shanghai, 200092 People's Republic of China.

(3)School of Information Engineering, Shanghai Maritime University, Shanghai,

201306 People's Republic of China. (4)Digestive Diseases and Nutrition Center,

Department of Pediatrics, The State University of New York at Buffalo, Buffalo,

NY 14260 USA ; Genomics, Environment, and Microbiome Community of Excellence, The

State University of New York at Buffalo, Buffalo, NY 14203 USA ; Institute of

Digestive Diseases, Longhua Hospital, Shanghai University of Traditional Chinese

Medicine, Shanghai, 200032 People's Republic of China.

MOTIVATION: Protein-binding sites prediction lays a foundation for functional

annotation of protein and structure-based drug design. As the number of available

protein structures increases, structural alignment based algorithm becomes the

dominant approach for protein-binding sites prediction. However, the present

algorithms underutilize the ever increasing numbers of three-dimensional

protein-ligand complex structures (bound protein), and it could be improved on

the process of alignment, selection of templates and clustering of template.

Herein, we built so far the largest database of bound templates with stringent

quality control. And on this basis, bSiteFinder as a protein-binding sites

prediction server was developed.

RESULTS: By introducing Homology Indexing, Chain Length Indexing, Stability of

Complex and Optimized Multiple-Templates Clustering into our algorithm, the

efficiency of our server has been significantly improved. Further, the accuracy

was approximately 2-10 % higher than that of other algorithms for the test with

either bound dataset or unbound dataset. For 210 bound dataset, bSiteFinder

achieved high accuracies up to 94.8 % (MCC 0.95). For another 48 bound/unbound

dataset, bSiteFinder achieved high accuracies up to 93.8 % for bound proteins

(MCC 0.95) and 85.4 % for unbound proteins (MCC 0.72). Our bSiteFinder server is

freely available at http://binfo.shmtu.edu.cn/bsitefinder/, and the source code

is provided at the methods page.

CONCLUSION: An online bSiteFinder server is freely available at

http://binfo.shmtu.edu.cn/bsitefinder/. Our work lays a foundation for functional

annotation of protein and structure-based drug design. With ever increasing

numbers of three-dimensional protein-ligand complex structures, our server should

be more accurate and less time-consuming.Graphical Abstract bSiteFinder

(http://binfo.shmtu.edu.cn/bsitefinder/) as a protein-binding sites prediction

server was developed based on the largest database of bound templates so far with

stringent quality control. By introducing Homology Indexing, Chain Length

Indexing, Stability of Complex and Optimized Multiple-Templates Clustering into

our algorithm, the efficiency of our server have been significantly improved.

What's more, the accuracy was approximately 2-10 % higher than that of other

algorithms for the test with either bound dataset or unbound dataset.

DOI: 10.1186/s13321-016-0149-z

PMCID: PMC4939519

PMID: 27403208

150. Nucleic Acids Res. 2016 Jul 8;44(W1):W181-4. doi: 10.1093/nar/gkw459. Epub 2016

May 29.

miRNAFold: a web server for fast miRNA precursor prediction in genomes.

Tav C(1), Tempel S(1), Poligny L(1), Tahi F(2).

Author information:

(1)IBISC-IBGBI, Université Evry, Genopole, 23 Boulevard de France, 91037 Evry

CEDEX, France. (2)IBISC-IBGBI, Université Evry, Genopole, 23 Boulevard de France,

91037 Evry CEDEX, France Institute of Plant Sciences Paris Saclay (IPS2), CNRS,

INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay, Batiment

630, 91405 Orsay, France fariza.tahi@ibisc.univ-evry.fr.

Computational methods are required for prediction of non-coding RNAs (ncRNAs),

which are involved in many biological processes, especially at

post-transcriptional level. Among these ncRNAs, miRNAs have been largely studied

and biologists need efficient and fast tools for their identification. In

particular, ab initio methods are usually required when predicting novel miRNAs.

Here we present a web server dedicated for miRNA precursors identification at a

large scale in genomes. It is based on an algorithm called miRNAFold that allows

predicting miRNA hairpin structures quickly with high sensitivity. miRNAFold is

implemented as a web server with an intuitive and user-friendly interface, as

well as a standalone version. The web server is freely available at:

http://EvryRNA.ibisc.univ-evry.fr/miRNAFold.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw459

PMCID: PMC4987958

PMID: 27242364

151. Nucleic Acids Res. 2016 Jul 8;44(W1):W469-73. doi: 10.1093/nar/gkw458. Epub 2016

May 23.

mCSM-AB: a web server for predicting antibody-antigen affinity changes upon

mutation with graph-based signatures.

Pires DE(1), Ascher DB(2).

Author information:

(1)Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Belo Horizonte

30190-002, Brazil douglas.pires@cpqrr.fiocruz.br. (2)Centro de Pesquisas René

Rachou, Fundação Oswaldo Cruz, Belo Horizonte 30190-002, Brazil Department of

Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK Department of

Biochemistry, University of Melbourne, Victoria 3010, Australia

dascher@svi.edu.au.

Computational methods have traditionally struggled to predict the effect of

mutations in antibody-antigen complexes on binding affinity. This has limited

their usefulness during antibody engineering and development, and their ability

to predict biologically relevant escape mutations. Here we present mCSM-AB, a

user-friendly web server for accurately predicting antibody-antigen affinity

changes upon mutation which relies on graph-based signatures. We show that

mCSM-AB performs better than comparable methods that have been previously used

for antibody engineering. mCSM-AB web server is available at

http://structure.bioc.cam.ac.uk/mcsm\_ab.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw458

PMCID: PMC4987957

PMID: 27216816

152. Nucleic Acids Res. 2016 Jul 8;44(W1):W128-34. doi: 10.1093/nar/gkw455. Epub 2016

May 20.

DIANA-mirExTra v2.0: Uncovering microRNAs and transcription factors with crucial

roles in NGS expression data.

Vlachos IS(1), Vergoulis T(2), Paraskevopoulou MD(3), Lykokanellos F(2),

Georgakilas G(3), Georgiou P(2), Chatzopoulos S(2), Karagkouni D(3),

Christodoulou F(4), Dalamagas T(2), Hatzigeorgiou AG(5).

Author information:

(1)DIANA-Lab, Department of Electrical & Computer Engineering, University of

Thessaly, 38221 Volos, Greece Hellenic Pasteur Institute, 127 Vasilissis Sofias

Avenue, 11521 Athens, Greece Laboratory for Experimental Surgery and Surgical

Research 'N.S. Christeas', Medical School of Athens, University of Athens, Athens

11527, Greece ivlachos@lessr.eu. (2)'Athena' Research and Innovation Center,

11524 Athens, Greece. (3)DIANA-Lab, Department of Electrical & Computer

Engineering, University of Thessaly, 38221 Volos, Greece Hellenic Pasteur

Institute, 127 Vasilissis Sofias Avenue, 11521 Athens, Greece. (4)Miroculus, San

Francisco, CA 94107, USA. (5)DIANA-Lab, Department of Electrical & Computer

Engineering, University of Thessaly, 38221 Volos, Greece Hellenic Pasteur

Institute, 127 Vasilissis Sofias Avenue, 11521 Athens, Greece arhatzig@uth.gr.

Differential expression analysis (DEA) is one of the main instruments utilized

for revealing molecular mechanisms in pathological and physiological conditions.

DIANA-mirExTra v2.0 (http://www.microrna.gr/mirextrav2) performs a combined DEA

of mRNAs and microRNAs (miRNAs) to uncover miRNAs and transcription factors (TFs)

playing important regulatory roles between two investigated states. The web

server uses as input miRNA/RNA-Seq read count data sets that can be uploaded for

analysis. Users can combine their data with 350 small-RNA-Seq and 65 RNA-Seq

in-house analyzed libraries which are provided by DIANA-mirExTra v2.0.The web

server utilizes miRNA:mRNA, TF:mRNA and TF:miRNA interactions derived from

extensive experimental data sets. More than 450 000 miRNA interactions and 2 000

000 TF binding sites from specific or high-throughput techniques have been

incorporated, while accurate miRNA TSS annotation is obtained from microTSS

experimental/in silico framework. These comprehensive data sets enable users to

perform analyses based solely on experimentally supported information and to

uncover central regulators within sequencing data: miRNAs controlling mRNAs and

TFs regulating mRNA or miRNA expression. The server also supports predicted

miRNA:gene interactions from DIANA-microT-CDS for 4 species (human, mouse,

nematode and fruit fly). DIANA-mirExTra v2.0 has an intuitive user interface and

is freely available to all users without any login requirement.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw455

PMCID: PMC4987956

PMID: 27207881

153. Nucleic Acids Res. 2016 Jul 8;44(W1):W568-74. doi: 10.1093/nar/gkw454. Epub 2016

May 19.

BindUP: a web server for non-homology-based prediction of DNA and RNA binding

proteins.

Paz I(1), Kligun E(1), Bengad B(1), Mandel-Gutfreund Y(2).

Author information:

(1)Department of Biology, Technion-Israel Institute of Technology, Technion City,

Haifa 32000, Israel. (2)Department of Biology, Technion-Israel Institute of

Technology, Technion City, Haifa 32000, Israel yaelmg@tx.technion.ac.il.

Gene expression is a multi-step process involving many layers of regulation. The

main regulators of the pathway are DNA and RNA binding proteins. While over the

years, a large number of DNA and RNA binding proteins have been identified and

extensively studied, it is still expected that many other proteins, some with yet

another known function, are awaiting to be discovered. Here we present a new web

server, BindUP, freely accessible through the website

http://bindup.technion.ac.il/, for predicting DNA and RNA binding proteins using

a non-homology-based approach. Our method is based on the electrostatic features

of the protein surface and other general properties of the protein. BindUP

predicts nucleic acid binding function given the proteins three-dimensional

structure or a structural model. Additionally, BindUP provides information on the

largest electrostatic surface patches, visualized on the server. The server was

tested on several datasets of DNA and RNA binding proteins, including proteins

which do not possess DNA or RNA binding domains and have no similarity to known

nucleic acid binding proteins, achieving very high accuracy. BindUP is applicable

in either single or batch modes and can be applied for testing hundreds of

proteins simultaneously in a highly efficient manner.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw454

PMCID: PMC4987955

PMID: 27198220

154. Nucleic Acids Res. 2016 Jul 8;44(W1):W367-74. doi: 10.1093/nar/gkw315. Epub 2016

May 19.

The RING 2.0 web server for high quality residue interaction networks.

Piovesan D(1), Minervini G(1), Tosatto SC(2).

Author information:

(1)Department of Biomedical Sciences, University of Padua, Padua 35121, Italy.

(2)Department of Biomedical Sciences, University of Padua, Padua 35121, Italy CNR

Institute of Neuroscience, Padua 35121, Italy silvio.tosatto@unipd.it.

Residue interaction networks (RINs) are an alternative way of representing

protein structures where nodes are residues and arcs physico-chemical

interactions. RINs have been extensively and successfully used for analysing

mutation effects, protein folding, domain-domain communication and catalytic

activity. Here we present RING 2.0, a new version of the RING software for the

identification of covalent and non-covalent bonds in protein structures,

including π-π stacking and π-cation interactions. RING 2.0 is extremely fast and

generates both intra and inter-chain interactions including solvent and ligand

atoms. The generated networks are very accurate and reliable thanks to a complex

empirical re-parameterization of distance thresholds performed on the entire

Protein Data Bank. By default, RING output is generated with optimal parameters

but the web server provides an exhaustive interface to customize the calculation.

The network can be visualized directly in the browser or in Cytoscape.

Alternatively, the RING-Viz script for Pymol allows visualizing the interactions

at atomic level in the structure. The web server and RING-Viz, together with an

extensive help and tutorial, are available from URL:

http://protein.bio.unipd.it/ring.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw315

PMCID: PMC4987896

PMID: 27198219

155. Nucleic Acids Res. 2016 Jul 8;44(W1):W147-53. doi: 10.1093/nar/gkw419. Epub 2016

May 17.

Heatmapper: web-enabled heat mapping for all.

Babicki S(1), Arndt D(1), Marcu A(1), Liang Y(1), Grant JR(1), Maciejewski A(2),

Wishart DS(3).

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8,

Canada. (2)Department of Biological Sciences, University of Alberta, Edmonton, AB

T6G 2E9, Canada. (3)Department of Computing Science, University of Alberta,

Edmonton, AB T6G 2E8, Canada Department of Biological Sciences, University of

Alberta, Edmonton, AB T6G 2E9, Canada National Institute for Nanotechnology,

11421 Saskatchewan Drive, Edmonton, AB T6G 2M9, Canada david.wishart@ualberta.ca.

Heatmapper is a freely available web server that allows users to interactively

visualize their data in the form of heat maps through an easy-to-use graphical

interface. Unlike existing non-commercial heat map packages, which either lack

graphical interfaces or are specialized for only one or two kinds of heat maps,

Heatmapper is a versatile tool that allows users to easily create a wide variety

of heat maps for many different data types and applications. More specifically,

Heatmapper allows users to generate, cluster and visualize: (i) expression-based

heat maps from transcriptomic, proteomic and metabolomic experiments; (ii)

pairwise distance maps; (iii) correlation maps; (iv) image overlay heat maps; (v)

latitude and longitude heat maps and (vi) geopolitical (choropleth) heat maps.

Heatmapper offers a number of simple and intuitive customization options for

facile adjustments to each heat map's appearance and plotting parameters.

Heatmapper also allows users to interactively explore their numeric data values

by hovering their cursor over each heat map cell, or by using a

searchable/sortable data table view. Heat map data can be easily uploaded to

Heatmapper in text, Excel or tab delimited formatted tables and the resulting

heat map images can be easily downloaded in common formats including PNG, JPG and

PDF. Heatmapper is designed to appeal to a wide range of users, including

molecular biologists, structural biologists, microbiologists, epidemiologists,

environmental scientists, agriculture/forestry scientists, fish and wildlife

biologists, climatologists, geologists, educators and students. Heatmapper is

available at http://www.heatmapper.ca.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw419

PMCID: PMC4987948

PMID: 27190236

156. Nucleic Acids Res. 2016 Jul 8;44(W1):W328-32. doi: 10.1093/nar/gkw412. Epub 2016

May 16.

iPARTS2: an improved tool for pairwise alignment of RNA tertiary structures,

version 2.

Yang CH(1), Shih CT(2), Chen KT(2), Lee PH(2), Tsai PH(2), Lin JC(2), Yen CY(2),

Lin TY(3), Lu CL(4).

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu 30050, Taiwan Department of Biological Science and

Technology, National Chiao Tung University, Hsinchu 30050, Taiwan. (2)Department

of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan.

(3)Department of Biological Science and Technology, National Chiao Tung

University, Hsinchu 30050, Taiwan. (4)Department of Computer Science, National

Tsing Hua University, Hsinchu 30013, Taiwan cllu@cs.nthu.edu.tw.

Since its first release in 2010, iPARTS has become a valuable tool for globally

or locally aligning two RNA 3D structures. It was implemented by a structural

alphabet (SA)-based approach, which uses an SA of 23 letters to reduce RNA 3D

structures into 1D sequences of SA letters and applies traditional sequence

alignment to these SA-encoded sequences for determining their global or local

similarity. In this version, we have re-implemented iPARTS into a new web server

iPARTS2 by constructing a totally new SA, which consists of 92 elements with each

carrying both information of base and backbone geometry for a representative

nucleotide. This SA is significantly different from the one used in iPARTS,

because the latter consists of only 23 elements with each carrying only the

backbone geometry information of a representative nucleotide. Our experimental

results have shown that iPARTS2 outperforms its previous version iPARTS and also

achieves better accuracy than other popular tools, such as SARA, SETTER and RASS,

in RNA alignment quality and function prediction. iPARTS2 takes as input two RNA

3D structures in the PDB format and outputs their global or local alignments with

graphical display. iPARTS2 is now available online at

http://genome.cs.nthu.edu.tw/iPARTS2/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw412

PMCID: PMC4987943

PMID: 27185896

157. Nucleic Acids Res. 2016 Jul 8;44(W1):W308-14. doi: 10.1093/nar/gkw440. Epub 2016

May 16.

incaRNAfbinv: a web server for the fragment-based design of RNA sequences.

Drory Retwitzer M(1), Reinharz V(2), Ponty Y(3), Waldispühl J(2), Barash D(4).

Author information:

(1)Department of Computer Science, Ben-Gurion University, Beer-Sheva 84105,

Israel. (2)School of Computer Science & McGill Centre for Bioinformatics, McGill

University, Montréal, QC H3A 0E9, Canada. (3)Laboratoire d'Informatique

(LIX)-CNRS UMR 7161, École Polytechnique, 91128 Palaiseau, France AMIB

team/project, INRIA Saclay, Bâtiment Alan Turing, 91128 Palaiseau, France.

(4)Department of Computer Science, Ben-Gurion University, Beer-Sheva 84105,

Israel dbarash@cs.bgu.ac.il.

In recent years, new methods for computational RNA design have been developed and

applied to various problems in synthetic biology and nanotechnology. Lately,

there is considerable interest in incorporating essential biological information

when solving the inverse RNA folding problem. Correspondingly, RNAfbinv aims at

including biologically meaningful constraints and is the only program to-date

that performs a fragment-based design of RNA sequences. In doing so it allows the

design of sequences that do not necessarily exactly fold into the target, as long

as the overall coarse-grained tree graph shape is preserved. Augmented by the

weighted sampling algorithm of incaRNAtion, our web server called incaRNAfbinv

implements the method devised in RNAfbinv and offers an interactive environment

for the inverse folding of RNA using a fragment-based design approach. It takes

as input: a target RNA secondary structure; optional sequence and motif

constraints; optional target minimum free energy, neutrality and GC content. In

addition to the design of synthetic regulatory sequences, it can be used as a

pre-processing step for the detection of novel natural occurring RNAs. The two

complementary methodologies RNAfbinv and incaRNAtion are merged together and

fully implemented in our web server incaRNAfbinv, available at

http://www.cs.bgu.ac.il/incaRNAfbinv.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw440

PMID: 27185893

158. Nucleic Acids Res. 2016 Jul 8;44(W1):W252-8. doi: 10.1093/nar/gkw434. Epub 2016

May 16.

CTLPScanner: a web server for chromothripsis-like pattern detection.

Yang J(1), Liu J(1), Ouyang L(2), Chen Y(3), Liu B(4), Cai H(5).

Author information:

(1)Center of Growth, Metabolism, and Aging, Key Laboratory of Bio-Resources and

Eco-Environment, College of Life Sciences, Sichuan University, Chengdu 610064,

Sichuan, China. (2)State Key Laboratory of Biotherapy, West China Hospital,

Sichuan University, and Collaborative Innovation Center of Biotherapy, Chengdu

610041, Sichuan, China. (3)State Key Laboratory of Biotherapy, West China

Hospital, Sichuan University, and Collaborative Innovation Center of Biotherapy,

Chengdu 610041, Sichuan, China Department of Gastrointestinal Surgery, West China

Hospital, Sichuan University, Chengdu 610041, Sichuan, China. (4)State Key

Laboratory of Biotherapy, West China Hospital, Sichuan University, and

Collaborative Innovation Center of Biotherapy, Chengdu 610041, Sichuan, China

liubo2400@scu.edu.cn. (5)Center of Growth, Metabolism, and Aging, Key Laboratory

of Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan

University, Chengdu 610064, Sichuan, China haoyang.cai@scu.edu.cn.

Chromothripsis is a recently observed phenomenon in cancer cells in which one or

several chromosomes shatter into pieces with subsequent inaccurate reassembly and

clonal propagation. This type of event generates a potentially vast number of

mutations within a relatively short-time period, and has been considered as a new

paradigm in cancer development. Despite recent advances, much work is still

required to better understand the molecular mechanisms of this phenomenon, and

thus an easy-to-use tool is in urgent need for automatically detecting and

annotating chromothripsis. Here we present CTLPScanner, a web server for

detection of chromothripsis-like pattern (CTLP) in genomic array data. The output

interface presents intuitive graphical representations of detected chromosome

pulverization region, as well as detailed results in table format. CTLPScanner

also provides additional information for associated genes in chromothripsis

region to help identify the potential candidates involved in tumorigenesis. To

assist in performing meta-data analysis, we integrated over 50 000 pre-processed

genomic arrays from The Cancer Genome Atlas and Gene Expression Omnibus into

CTLPScanner. The server allows users to explore the presence of chromothripsis

signatures from public data resources, without carrying out any local data

processing. CTLPScanner is freely available at

http://cgma.scu.edu.cn/CTLPScanner/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw434

PMCID: PMC4987951

PMID: 27185889

159. Nucleic Acids Res. 2016 Jul 8;44(W1):W185-93. doi: 10.1093/nar/gkw414. Epub 2016

May 13.

tRF2Cancer: A web server to detect tRNA-derived small RNA fragments (tRFs) and

their expression in multiple cancers.

Zheng LL(1), Xu WL(1), Liu S(1), Sun WJ(1), Li JH(1), Wu J(1), Yang JH(2), Qu

LH(3).

Author information:

(1)Key Laboratory of Gene Engineering of the Ministry of Education, State Key

Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University,

Guangzhou 510275, P. R. China. (2)Key Laboratory of Gene Engineering of the

Ministry of Education, State Key Laboratory of Biocontrol, School of Life

Sciences, Sun Yat-sen University, Guangzhou 510275, P. R. China

yangjh7@mail.sysu.edu.cn. (3)Key Laboratory of Gene Engineering of the Ministry

of Education, State Key Laboratory of Biocontrol, School of Life Sciences, Sun

Yat-sen University, Guangzhou 510275, P. R. China lssqlh@mail.sysu.edu.cn.

tRNA-derived small RNA fragments (tRFs) are one class of small non-coding RNAs

derived from transfer RNAs (tRNAs). tRFs play important roles in cellular

processes and are involved in multiple cancers. High-throughput small RNA (sRNA)

sequencing experiments can detect all the cellular expressed sRNAs, including

tRFs. However, distinguishing genuine tRFs from RNA fragments generated by random

degradation remains a major challenge. In this study, we developed an integrated

web-based computing system, tRF2Cancer, to accurately identify tRFs from sRNA

deep-sequencing data and evaluate their expression in multiple cancers. The

binomial test was introduced to evaluate whether reads from a small RNA-seq data

set represent tRFs or degraded fragments. A classification method was then used

to annotate the types of tRFs based on their sites of origin in pre-tRNA or

mature tRNA. We applied the pipeline to analyze 10 991 data sets from 32 types of

cancers and identified thousands of expressed tRFs. A tool called 'tRFinCancer'

was developed to facilitate the users to inspect the expression of tRFs across

different types of cancers. Another tool called 'tRFBrowser' shows both the sites

of origin and the distribution of chemical modification sites in tRFs on their

source tRNA. The tRF2Cancer web server is available at

http://rna.sysu.edu.cn/tRFfinder/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw414

PMCID: PMC4987945

PMID: 27179031

160. Nucleic Acids Res. 2016 Jul 8;44(W1):W166-75. doi: 10.1093/nar/gkw427. Epub 2016

May 13.

DeAnnIso: a tool for online detection and annotation of isomiRs from small RNA

sequencing data.

Zhang Y(1), Zang Q(2), Zhang H(3), Ban R(2), Yang Y(4), Iqbal F(5), Li A(6), Shi

Q(7).

Author information:

(1)Molecular and Cell Genetics Laboratory, The CAS Key Laboratory of Innate

Immunity and Chronic Disease, Hefei National Laboratory for Physical Sciences at

Microscale and School of Life Sciences, University of Science and Technology of

China, Hefei 230027, China Hefei Institute of Physical Science, China Academy of

Science, Hefei 230027, China. (2)School of Information Science and Technology,

University of Science and Technology of China, Hefei 230027, China. (3)Molecular

and Cell Genetics Laboratory, The CAS Key Laboratory of Innate Immunity and

Chronic Disease, Hefei National Laboratory for Physical Sciences at Microscale

and School of Life Sciences, University of Science and Technology of China, Hefei

230027, China. (4)Department of statistics, University of Kentucky, Lexington, KY

40536, USA. (5)Molecular and Cell Genetics Laboratory, The CAS Key Laboratory of

Innate Immunity and Chronic Disease, Hefei National Laboratory for Physical

Sciences at Microscale and School of Life Sciences, University of Science and

Technology of China, Hefei 230027, China Institute of Pure and Applied Biology,

Bahauddin Zakariya University, Multan 60800, Pakistan. (6)School of Information

Science and Technology, University of Science and Technology of China, Hefei

230027, China Research Centers for Biomedical Engineering, University of Science

and Technology of China, Hefei 230027, China aoli@ustc.edu.cn. (7)Molecular and

Cell Genetics Laboratory, The CAS Key Laboratory of Innate Immunity and Chronic

Disease, Hefei National Laboratory for Physical Sciences at Microscale and School

of Life Sciences, University of Science and Technology of China, Hefei 230027,

China Hefei Institute of Physical Science, China Academy of Science, Hefei

230027, China qshi@ustc.edu.cn.

Small RNA (sRNA) Sequencing technology has revealed that microRNAs (miRNAs) are

capable of exhibiting frequent variations from their canonical sequences,

generating multiple variants: the isoforms of miRNAs (isomiRs). However,

integrated tool to precisely detect and systematically annotate isomiRs from sRNA

sequencing data is still in great demand. Here, we present an online tool,

DeAnnIso (Detection and Annotation of IsomiRs from sRNA sequencing data).

DeAnnIso can detect all the isomiRs in an uploaded sample, and can extract the

differentially expressing isomiRs from paired or multiple samples. Once the

isomiRs detection is accomplished, detailed annotation information, including

isomiRs expression, isomiRs classification, SNPs in miRNAs and tissue specific

isomiR expression are provided to users. Furthermore, DeAnnIso provides a

comprehensive module of target analysis and enrichment analysis for the selected

isomiRs. Taken together, DeAnnIso is convenient for users to screen for isomiRs

of their interest and useful for further functional studies. The server is

implemented in PHP + Perl + R and available to all users for free at:

http://mcg.ustc.edu.cn/bsc/deanniso/ and http://mcg2.ustc.edu.cn/bsc/deanniso/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw427

PMCID: PMC4987950

PMID: 27179030

161. Nucleic Acids Res. 2016 Jul 8;44(W1):W205-11. doi: 10.1093/nar/gkw415. Epub 2016

May 12.

PASMet: a web-based platform for prediction, modelling and analyses of metabolic

systems.

Sriyudthsak K(1), Mejia RF(1), Arita M(2), Hirai MY(3).

Author information:

(1)RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa 230-0045,

Japan. (2)RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa

230-0045, Japan National Institute of Genetics, Mishima, Shizuoka 411-8540,

Japan. (3)RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa

230-0045, Japan masami.hirai@riken.jp.

PASMet (Prediction, Analysis and Simulation of Metabolic networks) is a web-based

platform for proposing and verifying mathematical models to understand the

dynamics of metabolism. The advantages of PASMet include user-friendliness and

accessibility, which enable biologists and biochemists to easily perform

mathematical modelling. PASMet offers a series of user-functions to handle the

time-series data of metabolite concentrations. The functions are organised into

four steps: (i) Prediction of a probable metabolic pathway and its regulation;

(ii) Construction of mathematical models; (iii) Simulation of metabolic

behaviours; and (iv) Analysis of metabolic system characteristics. Each function

contains various statistical and mathematical methods that can be used

independently. Users who may not have enough knowledge of computing or

programming can easily and quickly analyse their local data without software

downloads, updates or installations. Users only need to upload their files in

comma-separated values (CSV) format or enter their model equations directly into

the website. Once the time-series data or mathematical equations are uploaded,

PASMet automatically performs computation on server-side. Then, users can

interactively view their results and directly download them to their local

computers. PASMet is freely available with no login requirement at

http://pasmet.riken.jp/ from major web browsers on Windows, Mac and Linux

operating systems.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw415

PMCID: PMC4987946

PMID: 27174940

162. Nucleic Acids Res. 2016 Jul 8;44(W1):W54-7. doi: 10.1093/nar/gkw413. Epub 2016

May 12.

tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA

genes.

Lowe TM(1), Chan PP(2).

Author information:

(1)Department of Biomolecular Engineering, University of California Santa Cruz,

CA 95064, USA lowe@soe.ucsc.edu. (2)Department of Biomolecular Engineering,

University of California Santa Cruz, CA 95064, USA.

High-throughput genome sequencing continues to grow the need for rapid, accurate

genome annotation and tRNA genes constitute the largest family of essential,

ever-present non-coding RNA genes. Newly developed tRNAscan-SE 2.0 has advanced

the state-of-the-art methodology in tRNA gene detection and functional

prediction, captured by rich new content of the companion Genomic tRNA Database.

Previously, web-server tRNA detection was isolated from knowledge of existing

tRNAs and their annotation. In this update of the tRNAscan-SE On-line resource,

we tie together improvements in tRNA classification with greatly enhanced

biological context via dynamically generated links between web server search

results, the most relevant genes in the GtRNAdb and interactive, rich genome

context provided by UCSC genome browsers. The tRNAscan-SE On-line web server can

be accessed at http://trna.ucsc.edu/tRNAscan-SE/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw413

PMCID: PMC4987944

PMID: 27174935

163. Nucleic Acids Res. 2016 Jul 8;44(W1):W479-87. doi: 10.1093/nar/gkw416. Epub 2016

May 12.

HotSpot Wizard 2.0: automated design of site-specific mutations and smart

libraries in protein engineering.

Bendl J(1), Stourac J(2), Sebestova E(2), Vavra O(2), Musil M(3), Brezovsky J(4),

Damborsky J(5).

Author information:

(1)Loschmidt Laboratories, Department of Experimental Biology and Research Centre

for Toxic Compounds in the Environment RECETOX, Masaryk University, 625 00 Brno,

Czech Republic Department of Information Systems, Faculty of Information

Technology, Brno University of Technology, 612 66 Brno, Czech Republic

International Clinical Research Center, St. Anne's University Hospital Brno,

Brno, Czech Republic. (2)Loschmidt Laboratories, Department of Experimental

Biology and Research Centre for Toxic Compounds in the Environment RECETOX,

Masaryk University, 625 00 Brno, Czech Republic. (3)Loschmidt Laboratories,

Department of Experimental Biology and Research Centre for Toxic Compounds in the

Environment RECETOX, Masaryk University, 625 00 Brno, Czech Republic Department

of Information Systems, Faculty of Information Technology, Brno University of

Technology, 612 66 Brno, Czech Republic. (4)Loschmidt Laboratories, Department of

Experimental Biology and Research Centre for Toxic Compounds in the Environment

RECETOX, Masaryk University, 625 00 Brno, Czech Republic International Clinical

Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic

brezovsky@mail.muni.cz. (5)Loschmidt Laboratories, Department of Experimental

Biology and Research Centre for Toxic Compounds in the Environment RECETOX,

Masaryk University, 625 00 Brno, Czech Republic International Clinical Research

Center, St. Anne's University Hospital Brno, Brno, Czech Republic

jiri@chemi.muni.cz.

HotSpot Wizard 2.0 is a web server for automated identification of hot spots and

design of smart libraries for engineering proteins' stability, catalytic

activity, substrate specificity and enantioselectivity. The server integrates

sequence, structural and evolutionary information obtained from 3 databases and

20 computational tools. Users are guided through the processes of selecting hot

spots using four different protein engineering strategies and optimizing the

resulting library's size by narrowing down a set of substitutions at individual

randomized positions. The only required input is a query protein structure. The

results of the calculations are mapped onto the protein's structure and

visualized with a JSmol applet. HotSpot Wizard lists annotated residues suitable

for mutagenesis and can automatically design appropriate codons for each

implemented strategy. Overall, HotSpot Wizard provides comprehensive annotations

of protein structures and assists protein engineers with the rational design of

site-specific mutations and focused libraries. It is freely available at

http://loschmidt.chemi.muni.cz/hotspotwizard.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw416

PMCID: PMC4987947

PMID: 27174934

164. Nucleic Acids Res. 2016 Jul 8;44(W1):W488-93. doi: 10.1093/nar/gkw409. Epub 2016

May 12.

MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein

sequences.

Malhis N(1), Jacobson M(2), Gsponer J(3).

Author information:

(1)Michael Smith Laboratories-Centre for High-Throughput Biology, The University

of British Columbia, Vancouver, BC V6T 1Z4, Canada nmalhis@chibi.ubc.ca.

(2)Michael Smith Laboratories-Centre for High-Throughput Biology, The University

of British Columbia, Vancouver, BC V6T 1Z4, Canada. (3)Michael Smith

Laboratories-Centre for High-Throughput Biology, The University of British

Columbia, Vancouver, BC V6T 1Z4, Canada Department of Biochemistry and Molecular

Biology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

gsponer@chibi.ubc.ca.

Molecular recognition features, MoRFs, are short segments within longer

disordered protein regions that bind to globular protein domains in a process

known as disorder-to-order transition. MoRFs have been found to play a

significant role in signaling and regulatory processes in cells. High-confidence

computational identification of MoRFs remains an important challenge. In this

work, we introduce MoRFchibi SYSTEM that contains three MoRF predictors:

MoRFCHiBi, a basic predictor best suited as a component in other applications,

MoRFCHiBi\_ Light, ideal for high-throughput predictions and MoRFCHiBi\_ Web,

slower than the other two but best for high accuracy predictions. Results show

that MoRFchibi SYSTEM provides more than double the precision of other

predictors. MoRFchibi SYSTEM is available in three different forms: as HTML web

server, RESTful web server and downloadable software at:

http://www.chibi.ubc.ca/faculty/joerg-gsponer/gsponer-lab/software/morf\_chibi/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw409

PMCID: PMC4987941

PMID: 27174932

165. Nucleic Acids Res. 2016 Jul 8;44(W1):W333-8. doi: 10.1093/nar/gkw410. Epub 2016

May 12.

rMAPS: RNA map analysis and plotting server for alternative exon regulation.

Park JW(1), Jung S(2), Rouchka EC(3), Tseng YT(4), Xing Y(5).

Author information:

(1)Department of Computer Engineering and Computer Science, University of

Louisville, Louisville, KY 40292, USA KBRIN Bioinformatics Core, University of

Louisville, Louisville, KY 40202, USA Department of Microbiology, Immunology &

Molecular Genetics, University of California, Los Angeles, CA 90095, USA

juw.park@louisville.edu. (2)Department of Computer Engineering and Computer

Science, University of Louisville, Louisville, KY 40292, USA. (3)Department of

Computer Engineering and Computer Science, University of Louisville, Louisville,

KY 40292, USA KBRIN Bioinformatics Core, University of Louisville, Louisville, KY

40202, USA. (4)Department of Microbiology, Immunology & Molecular Genetics,

University of California, Los Angeles, CA 90095, USA. (5)Department of

Microbiology, Immunology & Molecular Genetics, University of California, Los

Angeles, CA 90095, USA yxing@ucla.edu.

RNA-binding proteins (RBPs) play a critical role in the regulation of alternative

splicing (AS), a prevalent mechanism for generating transcriptomic and proteomic

diversity in eukaryotic cells. Studies have shown that AS can be regulated by

RBPs in a binding-site-position dependent manner. Depending on where RBPs bind,

splicing of an alternative exon can be enhanced or suppressed. Therefore, spatial

analyses of RBP motifs and binding sites around alternative exons will help

elucidate splicing regulation by RBPs. The development of high-throughput

sequencing technologies has allowed transcriptome-wide analyses of AS and RBP-RNA

interactions. Given a set of differentially regulated alternative exons obtained

from RNA sequencing (RNA-seq) experiments, the rMAPS web server

(http://rmaps.cecsresearch.org) performs motif analyses of RBPs in the vicinity

of alternatively spliced exons and creates RNA maps that depict the spatial

patterns of RBP motifs. Similarly, rMAPS can also perform spatial analyses of

RBP-RNA binding sites identified by cross-linking immunoprecipitation sequencing

(CLIP-seq) experiments. We anticipate rMAPS will be a useful tool for elucidating

RBP regulation of alternative exon splicing using high-throughput sequencing

data.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw410

PMCID: PMC4987942

PMID: 27174931

166. Nucleic Acids Res. 2016 Jul 8;44(W1):W401-5. doi: 10.1093/nar/gkw332. Epub 2016

May 12.

PSSweb: protein structural statistics web server.

Gaillard T(1), Stote RH(2), Dejaegere A(2).

Author information:

(1)Laboratoire de Biochimie, École Polytechnique, CNRS, Université Paris-Saclay,

91128 Palaiseau cedex, France thomas.gaillard@polytechnique.edu. (2)Department of

Integrative Structural Biology, Institut de Génétique et de Biologie Moléculaire

et Cellulaire, Institut National de la Santé et de la Recherche Médicale U964,

Centre National de la Recherche Scientifique UMR 7104, Université de Strasbourg,

67404 Illkirch, France.

With the increasing number of protein structures available, there is a need for

tools capable of automating the comparison of ensembles of structures, a common

requirement in structural biology and bioinformatics. PSSweb is a web server for

protein structural statistics. It takes as input an ensemble of PDB files of

protein structures, performs a multiple sequence alignment and computes

structural statistics for each position of the alignment. Different optional

functionalities are proposed: structure superposition, Cartesian coordinate

statistics, dihedral angle calculation and statistics, and a cluster analysis

based on dihedral angles. An interactive report is generated, containing a

summary of the results, tables, figures and 3D visualization of superposed

structures. The server is available at http://pssweb.org.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw332

PMCID: PMC4987900

PMID: 27174930

167. Nucleic Acids Res. 2016 Jul 8;44(W1):W344-50. doi: 10.1093/nar/gkw408. Epub 2016

May 10.

ConSurf 2016: an improved methodology to estimate and visualize evolutionary

conservation in macromolecules.

Ashkenazy H(1), Abadi S(2), Martz E(3), Chay O(4), Mayrose I(5), Pupko T(6),

Ben-Tal N(7).

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, Tel Aviv 69978, Israel. (2)Department of Molecular

Biology and Ecology of Plants, George S. Wise Faculty of Life Sciences, Tel Aviv

University, Tel Aviv 69978, Israel. (3)Department of Microbiology, University of

Massachusetts, Amherst, MA 01003, USA. (4)Department of Cell Research and

Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel

Aviv 69978, Israel Department of Molecular Biology and Ecology of Plants, George

S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, Tel Aviv 69978, Israel. (5)Department of Molecular

Biology and Ecology of Plants, George S. Wise Faculty of Life Sciences, Tel Aviv

University, Tel Aviv 69978, Israel itaymay@post.tau.ac.il. (6)Department of Cell

Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv

University, Tel Aviv 69978, Israel talp@post.tau.ac.il. (7)Department of

Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel

Aviv University, Tel Aviv 69978, Israel bental@tauex.tau.ac.il.

The degree of evolutionary conservation of an amino acid in a protein or a

nucleic acid in DNA/RNA reflects a balance between its natural tendency to mutate

and the overall need to retain the structural integrity and function of the

macromolecule. The ConSurf web server (http://consurf.tau.ac.il), established

over 15 years ago, analyses the evolutionary pattern of the amino/nucleic acids

of the macromolecule to reveal regions that are important for structure and/or

function. Starting from a query sequence or structure, the server automatically

collects homologues, infers their multiple sequence alignment and reconstructs a

phylogenetic tree that reflects their evolutionary relations. These data are then

used, within a probabilistic framework, to estimate the evolutionary rates of

each sequence position. Here we introduce several new features into ConSurf,

including automatic selection of the best evolutionary model used to infer the

rates, the ability to homology-model query proteins, prediction of the secondary

structure of query RNA molecules from sequence, the ability to view the

biological assembly of a query (in addition to the single chain), mapping of the

conservation grades onto 2D RNA models and an advanced view of the phylogenetic

tree that enables interactively rerunning ConSurf with the taxa of a sub-tree.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw408

PMCID: PMC4987940

PMID: 27166375

168. Nucleic Acids Res. 2016 Jul 8;44(W1):W455-62. doi: 10.1093/nar/gkw403. Epub 2016

May 10.

GPCR-ModSim: A comprehensive web based solution for modeling G-protein coupled

receptors.

Esguerra M(1), Siretskiy A(1), Bello X(2), Sallander J(1), Gutiérrez-de-Terán

H(3).

Author information:

(1)Department of Cell and Molecular Biology, Uppsala University, Biomedical

Center, Box 596, SE-751 24, Uppsala, Sweden. (2)Fundación Pública Galega de

Medicina Xenómica, Hospital Clínico Universitario de Santiago, Santiago de

Compostela, 15706, Spain. (3)Department of Cell and Molecular Biology, Uppsala

University, Biomedical Center, Box 596, SE-751 24, Uppsala, Sweden

hugo.gutierrez@icm.uu.se.

GPCR-ModSim (http://open.gpcr-modsim.org) is a centralized and easy to use

service dedicated to the structural modeling of G-protein Coupled Receptors

(GPCRs). 3D molecular models can be generated from amino acid sequence by

homology-modeling techniques, considering different receptor conformations.

GPCR-ModSim includes a membrane insertion and molecular dynamics (MD)

equilibration protocol, which can be used to refine the generated model or any

GPCR structure uploaded to the server, including if desired non-protein elements

such as orthosteric or allosteric ligands, structural waters or ions. We herein

revise the main characteristics of GPCR-ModSim and present new functionalities.

The templates used for homology modeling have been updated considering the latest

structural data, with separate profile structural alignments built for inactive,

partially-active and active groups of templates. We have also added the

possibility to perform multiple-template homology modeling in a unique and

flexible way. Finally, our new MD protocol considers a series of distance

restraints derived from a recently identified conserved network of helical

contacts, allowing for a smoother refinement of the generated models which is

particularly advised when there is low homology to the available templates. GPCR-

ModSim has been tested on the GPCR Dock 2013 competition with satisfactory

results.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw403

PMCID: PMC4987938

PMID: 27166369

169. Nucleic Acids Res. 2016 Jul 8;44(W1):W267-71. doi: 10.1093/nar/gkw407. Epub 2016

May 10.

Breaking-Cas-interactive design of guide RNAs for CRISPR-Cas experiments for

ENSEMBL genomes.

Oliveros JC(1), Franch M(1), Tabas-Madrid D(1), San-León D(1), Montoliu L(2),

Cubas P(1), Pazos F(3).

Author information:

(1)National Centre for Biotechnology (CNB-CSIC), c/ Darwin 3, 28049 Madrid,

Spain. (2)National Centre for Biotechnology (CNB-CSIC), c/ Darwin 3, 28049

Madrid, Spain CIBERER, ISCIII, Madrid, Spain. (3)National Centre for

Biotechnology (CNB-CSIC), c/ Darwin 3, 28049 Madrid, Spain pazos@cnb.csic.es.

The CRISPR/Cas technology is enabling targeted genome editing in multiple

organisms with unprecedented accuracy and specificity by using RNA-guided

nucleases. A critical point when planning a CRISPR/Cas experiment is the design

of the guide RNA (gRNA), which directs the nuclease and associated machinery to

the desired genomic location. This gRNA has to fulfil the requirements of the

nuclease and lack homology with other genome sites that could lead to off-target

effects. Here we introduce the Breaking-Cas system for the design of gRNAs for

CRISPR/Cas experiments, including those based in the Cas9 nuclease as well as

others recently introduced. The server has unique features not available in other

tools, including the possibility of using all eukaryotic genomes available in

ENSEMBL (currently around 700), placing variable PAM sequences at 5' or 3' and

setting the guide RNA length and the scores per nucleotides. It can be freely

accessed at: http://bioinfogp.cnb.csic.es/tools/breakingcas, and the code is

available upon request.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw407

PMCID: PMC4987939

PMID: 27166368

170. Nucleic Acids Res. 2016 Jul 8;44(W1):W46-53. doi: 10.1093/nar/gkw394. Epub 2016

May 6.

Comparative transcriptomics across the prokaryotic tree of life.

Cohen O(1), Doron S(2), Wurtzel O(3), Dar D(2), Edelheit S(2), Karunker I(2),

Mick E(4), Sorek R(5).

Author information:

(1)Department of Molecular Genetics, Weizmann Institute of Science, Rehovot

76100, Israel Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA.

(2)Department of Molecular Genetics, Weizmann Institute of Science, Rehovot

76100, Israel. (3)Department of Molecular Genetics, Weizmann Institute of

Science, Rehovot 76100, Israel Whitehead Institute for Biomedical Research,

Cambridge, MA 02142, USA. (4)Department of Molecular Genetics, Weizmann Institute

of Science, Rehovot 76100, Israel Department of Systems Biology, Harvard Medical

School, Boston, MA 02115, USA. (5)Department of Molecular Genetics, Weizmann

Institute of Science, Rehovot 76100, Israel rotem.sorek@weizmann.ac.il.

Whole-transcriptome sequencing studies from recent years revealed an unexpected

complexity in transcriptomes of bacteria and archaea, including abundant

non-coding RNAs, cis-antisense transcription and regulatory untranslated regions

(UTRs). Understanding the functional relevance of the plethora of non-coding RNAs

in a given organism is challenging, especially since some of these RNAs were

attributed to 'transcriptional noise'. To allow the search for conserved

transcriptomic elements we produced comparative transcriptome maps for multiple

species across the microbial tree of life. These transcriptome maps are detailed

in annotations, comparable by gene families, and BLAST-searchable by user

provided sequences. Our transcriptome collection includes 18 model organisms

spanning 10 phyla/subphyla of bacteria and archaea that were sequenced using

standardized RNA-seq methods. The utility of the comparative approach, as

implemented in our web server, is demonstrated by highlighting genes with

exceptionally long 5'UTRs across species, which correspond to many known

riboswitches and further suggest novel putative regulatory elements. Our study

provides a standardized reference transcriptome to major clinically and

environmentally important microbial phyla. The viewer is available at

http://exploration.weizmann.ac.il/TCOL, setting a framework for comparative

studies of the microbial non-coding genome.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw394

PMCID: PMC4987935

PMID: 27154273

171. Nucleic Acids Res. 2016 Jul 8;44(W1):W557-61. doi: 10.1093/nar/gkw390. Epub 2016

May 5.

CSM-lig: a web server for assessing and comparing protein-small molecule

affinities.

Pires DE(1), Ascher DB(2).

Author information:

(1)Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Belo Horizonte,

30190-002, Brazil douglas.pires@cpqrr.fiocruz.br. (2)Centro de Pesquisas René

Rachou, Fundação Oswaldo Cruz, Belo Horizonte, 30190-002, Brazil Department of

Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK Department of

Biochemistry, University of Melbourne, Victoria 3010, Australia

dascher@svi.edu.au.

Determining the affinity of a ligand for a given protein is a crucial component

of drug development and understanding their biological effects. Predicting

binding affinities is a challenging and difficult task, and despite being

regarded as poorly predictive, scoring functions play an important role in the

analysis of molecular docking results. Here, we present CSM-Lig

(http://structure.bioc.cam.ac.uk/csm\_lig), a web server tailored to predict the

binding affinity of a protein-small molecule complex, encompassing both protein

and small-molecule complementarity in terms of shape and chemistry via

graph-based structural signatures. CSM-Lig was trained and evaluated on different

releases of the PDBbind databases, achieving a correlation of up to 0.86 on

10-fold cross validation and 0.80 in blind tests, performing as well as or better

than other widely used methods. The web server allows users to rapidly and

automatically predict binding affinities of collections of structures and assess

the interactions made. We believe CSM-lig would be an invaluable tool for helping

assess docking poses, the effects of multiple mutations, including insertions,

deletions and alternative splicing events, in protein-small molecule affinity,

unraveling important aspects that drive protein-compound recognition.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw390

PMCID: PMC4987933

PMID: 27151202

172. Nucleic Acids Res. 2016 Jul 8;44(W1):W375-82. doi: 10.1093/nar/gkw383. Epub 2016

May 5.

NAPS: Network Analysis of Protein Structures.

Chakrabarty B(1), Parekh N(2).

Author information:

(1)Centre for Computational Natural Sciences and Bioinformatics, International

Institute of Information Technology, Hyderabad 500032, India. (2)Centre for

Computational Natural Sciences and Bioinformatics, International Institute of

Information Technology, Hyderabad 500032, India nita@iiit.ac.in.

Traditionally, protein structures have been analysed by the secondary structure

architecture and fold arrangement. An alternative approach that has shown promise

is modelling proteins as a network of non-covalent interactions between amino

acid residues. The network representation of proteins provide a systems approach

to topological analysis of complex three-dimensional structures irrespective of

secondary structure and fold type and provide insights into structure-function

relationship. We have developed a web server for network based analysis of

protein structures, NAPS, that facilitates quantitative and qualitative (visual)

analysis of residue-residue interactions in: single chains, protein complex,

modelled protein structures and trajectories (e.g. from molecular dynamics

simulations). The user can specify atom type for network construction, distance

range (in Å) and minimal amino acid separation along the sequence. NAPS provides

users selection of node(s) and its neighbourhood based on centrality measures,

physicochemical properties of amino acids or cluster of well-connected residues

(k-cliques) for further analysis. Visual analysis of interacting domains and

protein chains, and shortest path lengths between pair of residues are additional

features that aid in functional analysis. NAPS support various analyses and

visualization views for identifying functional residues, provide insight into

mechanisms of protein folding, domain-domain and protein-protein interactions for

understanding communication within and between proteins.

URL:http://bioinf.iiit.ac.in/NAPS/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw383

PMCID: PMC4987928

PMID: 27151201

173. Nucleic Acids Res. 2016 Jul 8;44(W1):W395-400. doi: 10.1093/nar/gkw395. Epub 2016

May 5.

RCD+: Fast loop modeling server.

López-Blanco JR(1), Canosa-Valls AJ(1), Li Y(2), Chacón P(3).

Author information:

(1)Department of Biological Chemical Physics, Rocasolano Physical Chemistry

Institute C.S.I.C., Serrano 119, 28006 Madrid, Spain. (2)Department of Computer

Science, Old Dominion University, Norfolk, VA 23529, USA. (3)Department of

Biological Chemical Physics, Rocasolano Physical Chemistry Institute C.S.I.C.,

Serrano 119, 28006 Madrid, Spain pablo@chaconlab.org.

Modeling loops is a critical and challenging step in protein modeling and

prediction. We have developed a quick online service (http://rcd.chaconlab.org)

for ab initio loop modeling combining a coarse-grained conformational search with

a full-atom refinement. Our original Random Coordinate Descent (RCD) loop closure

algorithm has been greatly improved to enrich the sampling distribution towards

near-native conformations. These improvements include a new workflow

optimization, MPI-parallelization and fast backbone angle sampling based on

neighbor-dependent Ramachandran probability distributions. The server starts by

efficiently searching the vast conformational space from only the loop sequence

information and the environment atomic coordinates. The generated closed loop

models are subsequently ranked using a fast distance-orientation dependent energy

filter. Top ranked loops are refined with the Rosetta energy function to obtain

accurate all-atom predictions that can be interactively inspected in an

user-friendly web interface. Using standard benchmarks, the average root mean

squared deviation (RMSD) is 0.8 and 1.4 Å for 8 and 12 residues loops,

respectively, in the challenging modeling scenario in where the side chains of

the loop environment are fully remodeled. These results are not only very

competitive compared to those obtained with public state of the art methods, but

also they are obtained ∼10-fold faster.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw395

PMCID: PMC4987936

PMID: 27151199

174. Nucleic Acids Res. 2016 Jul 8;44(W1):W105-9. doi: 10.1093/nar/gkw356. Epub 2016

May 5.

PathwAX: a web server for network crosstalk based pathway annotation.

Ogris C(1), Helleday T(2), Sonnhammer EL(3).

Author information:

(1)Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics,

Stockholm University, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden

christoph.ogris@scilifelab.se. (2)Division of Translational Medicine and Chemical

Biology, Department of Medical Biochemistry and Biophysics, Karolinska

Institutet, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden.

(3)Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics,

Stockholm University, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden

erik.sonnhammer@scilifelab.se.

Pathway annotation of gene lists is often used to functionally analyse

biomolecular data such as gene expression in order to establish which processes

are activated in a given experiment. Databases such as KEGG or GO represent

collections of how genes are known to be organized in pathways, and the challenge

is to compare a given gene list with the known pathways such that all true

relations are identified. Most tools apply statistical measures to the gene

overlap between the gene list and pathway. It is however problematic to avoid

false negatives and false positives when only using the gene overlap. The pathwAX

web server (http://pathwAX.sbc.su.se/) applies a different approach which is

based on network crosstalk. It uses the comprehensive network FunCoup to analyse

network crosstalk between a query gene list and KEGG pathways. PathwAX runs the

BinoX algorithm, which employs Monte-Carlo sampling of randomized networks and

estimates a binomial distribution, for estimating the statistical significance of

the crosstalk. This results in substantially higher accuracy than gene overlap

methods. The system was optimized for speed and allows interactive web usage. We

illustrate the usage and output of pathwAX.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw356

PMCID: PMC4987909

PMID: 27151197

175. Nucleic Acids Res. 2016 Jul 8;44(W1):W416-23. doi: 10.1093/nar/gkw368. Epub 2016

May 5.

BetaSCPWeb: side-chain prediction for protein structures using Voronoi diagrams

and geometry prioritization.

Ryu J(1), Lee M(1), Cha J(1), Laskowski RA(2), Ryu SE(3), Kim DS(4).

Author information:

(1)Vorononi Diagram Research Center, Hanyang University, Korea. (2)European

Bioinformatics Institute, Wellcome Trust Genome Campus, UK. (3)Department of

Bioengineering, Hanyang University, Korea. (4)School of Mechanical Engineering,

Hanyang University, Korea dskim@hanyang.ac.kr.

Many applications, such as protein design, homology modeling, flexible docking,

etc. require the prediction of a protein's optimal side-chain conformations from

just its amino acid sequence and backbone structure. Side-chain prediction (SCP)

is an NP-hard energy minimization problem. Here, we present BetaSCPWeb which

efficiently computes a conformation close to optimal using a

geometry-prioritization method based on the Voronoi diagram of spherical atoms.

Its outputs are visual, textual and PDB file format. The web server is free and

open to all users at http://voronoi.hanyang.ac.kr/betascpweb with no login

requirement.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw368

PMCID: PMC4987919

PMID: 27151195

176. Nucleic Acids Res. 2016 Jul 8;44(W1):W463-8. doi: 10.1093/nar/gkw364. Epub 2016

May 5.

StructMAn: annotation of single-nucleotide polymorphisms in the structural

context.

Gress A(1), Ramensky V(2), Büch J(3), Keller A(4), Kalinina OV(5).

Author information:

(1)Department for Computational Biology and Applied Algorithmics, Max Planck

Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany Graduate

School of Computer Science, Saarland University, Campus E1 3, 66123 Saarbrücken,

Germany. (2)Center for Neurobehavioral Genetics, University of California, Los

Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA.

(3)Department for Computational Biology and Applied Algorithmics, Max Planck

Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany. (4)Chair for

Medical Bioinformatics, Saarland University, Campus E2 2, 66123 Saarbrücken,

Germany. (5)Department for Computational Biology and Applied Algorithmics, Max

Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany

kalinina@mpi-inf.mpg.de.

The next generation sequencing technologies produce unprecedented amounts of data

on the genetic sequence of individual organisms. These sequences carry a

substantial amount of variation that may or may be not related to a phenotype.

Phenotypically important part of this variation often comes in form of

protein-sequence altering (non-synonymous) single nucleotide variants (nsSNVs).

Here we present StructMAn, a Web-based tool for annotation of human and non-human

nsSNVs in the structural context. StructMAn analyzes the spatial location of the

amino acid residue corresponding to nsSNVs in the three-dimensional (3D) protein

structure relative to other proteins, nucleic acids and low molecular-weight

ligands. We make use of all experimentally available 3D structures of query

proteins, and also, unlike other tools in the field, of structures of proteins

with detectable sequence identity to them. This allows us to provide a structural

context for around 20% of all nsSNVs in a typical human sequencing sample, for up

to 60% of nsSNVs in genes related to human diseases and for around 35% of nsSNVs

in a typical bacterial sample. Each nsSNV can be visualized and inspected by the

user in the corresponding 3D structure of a protein or protein complex. The

StructMAn server is available at http://structman.mpi-inf.mpg.de.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw364

PMCID: PMC4987916

PMID: 27150811

177. Nucleic Acids Res. 2016 Jul 8;44(W1):W494-501. doi: 10.1093/nar/gkw374. Epub 2016

May 5.

MutaBind estimates and interprets the effects of sequence variants on

protein-protein interactions.

Li M(1), Simonetti FL(2), Goncearenco A(1), Panchenko AR(3).

Author information:

(1)National Center for Biotechnology Information, National Institutes of Health,

Bethesda, MD 20894, USA. (2)Institute Leloir Foundation, Buenos Aires, C1405BWE,

Argentina. (3)National Center for Biotechnology Information, National Institutes

of Health, Bethesda, MD 20894, USA panch@ncbi.nlm.nih.gov.

Proteins engage in highly selective interactions with their macromolecular

partners. Sequence variants that alter protein binding affinity may cause

significant perturbations or complete abolishment of function, potentially

leading to diseases. There exists a persistent need to develop a mechanistic

understanding of impacts of variants on proteins. To address this need we

introduce a new computational method MutaBind to evaluate the effects of sequence

variants and disease mutations on protein interactions and calculate the

quantitative changes in binding affinity. The MutaBind method uses molecular

mechanics force fields, statistical potentials and fast side-chain optimization

algorithms. The MutaBind server maps mutations on a structural protein complex,

calculates the associated changes in binding affinity, determines the deleterious

effect of a mutation, estimates the confidence of this prediction and produces a

mutant structural model for download. MutaBind can be applied to a large number

of problems, including determination of potential driver mutations in cancer and

other diseases, elucidation of the effects of sequence variants on protein

fitness in evolution and protein design. MutaBind is available at

http://www.ncbi.nlm.nih.gov/projects/mutabind/.

Published by Oxford University Press on behalf of Nucleic Acids Research 2016.

This work is written by (a) US Government employee(s) and is in the public domain

in the US.

DOI: 10.1093/nar/gkw374

PMCID: PMC4987923

PMID: 27150810

178. Nucleic Acids Res. 2016 Jul 8;44(W1):W550-6. doi: 10.1093/nar/gkw393. Epub 2016

May 5.

ACFIS: a web server for fragment-based drug discovery.

Hao GF(1), Jiang W(1), Ye YN(2), Wu FX(1), Zhu XL(1), Guo FB(3), Yang GF(4).

Author information:

(1)Key Laboratory of Pesticide & Chemical Biology, Ministry of Education, College

of Chemistry, Central China Normal University, Wuhan 430079, P.R.China.

(2)Collaborative Innovation Center of Chemical Science and Engineering, Tianjing

300072, P.R.China. (3)Collaborative Innovation Center of Chemical Science and

Engineering, Tianjing 300072, P.R.China fbguo@uestc.edu.cn. (4)Key Laboratory of

Pesticide & Chemical Biology, Ministry of Education, College of Chemistry,

Central China Normal University, Wuhan 430079, P.R.China Collaborative Innovation

Center of Chemical Science and Engineering, Tianjing 300072, P.R.China

gfyang@mail.ccnu.edu.cn.

In order to foster innovation and improve the effectiveness of drug discovery,

there is a considerable interest in exploring unknown 'chemical space' to

identify new bioactive compounds with novel and diverse scaffolds. Hence,

fragment-based drug discovery (FBDD) was developed rapidly due to its advanced

expansive search for 'chemical space', which can lead to a higher hit rate and

ligand efficiency (LE). However, computational screening of fragments is always

hampered by the promiscuous binding model. In this study, we developed a new web

server Auto Core Fragment in silico Screening (ACFIS). It includes three

computational modules, PARA\_GEN, CORE\_GEN and CAND\_GEN. ACFIS can generate core

fragment structure from the active molecule using fragment deconstruction

analysis and perform in silico screening by growing fragments to the junction of

core fragment structure. An integrated energy calculation rapidly identifies

which fragments fit the binding site of a protein. We constructed a simple

interface to enable users to view top-ranking molecules in 2D and the binding

mode in 3D for further experimental exploration. This makes the ACFIS a highly

valuable tool for drug discovery. The ACFIS web server is free and open to all

users at http://chemyang.ccnu.edu.cn/ccb/server/ACFIS/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw393

PMCID: PMC4987934

PMID: 27150808

179. Nucleic Acids Res. 2016 Jul 8;44(W1):W16-21. doi: 10.1093/nar/gkw387. Epub 2016

May 3.

PHASTER: a better, faster version of the PHAST phage search tool.

Arndt D(1), Grant JR(1), Marcu A(1), Sajed T(1), Pon A(1), Liang Y(1), Wishart

DS(2).

Author information:

(1)Department of Computing Science, Edmonton, AB T6G 2E8, Canada. (2)Department

of Computing Science, Edmonton, AB T6G 2E8, Canada Department of Biological

Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada National Institute

for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, AB T6G 2M9, Canada

david.wishart@ualberta.ca.

PHASTER (PHAge Search Tool - Enhanced Release) is a significant upgrade to the

popular PHAST web server for the rapid identification and annotation of prophage

sequences within bacterial genomes and plasmids. Although the steps in the phage

identification pipeline in PHASTER remain largely the same as in the original

PHAST, numerous software improvements and significant hardware enhancements have

now made PHASTER faster, more efficient, more visually appealing and much more

user friendly. In particular, PHASTER is now 4.3× faster than PHAST when

analyzing a typical bacterial genome. More specifically, software optimizations

have made the backend of PHASTER 2.7X faster than PHAST, while the addition of 80

CPUs to the PHASTER compute cluster are responsible for the remaining speed-up.

PHASTER can now process a typical bacterial genome in 3 min from the raw sequence

alone, or in 1.5 min when given a pre-annotated GenBank file. A number of other

optimizations have also been implemented, including automated algorithms to

reduce the size and redundancy of PHASTER's databases, improvements in handling

multiple (metagenomic) queries and higher user traffic, along with the ability to

perform automated look-ups against 14 000 previously PHAST/PHASTER annotated

bacterial genomes (which can lead to complete phage annotations in seconds as

opposed to minutes). PHASTER's web interface has also been entirely rewritten. A

new graphical genome browser has been added, gene/genome visualization tools have

been improved, and the graphical interface is now more modern, robust and

user-friendly. PHASTER is available online at www.phaster.ca.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw387

PMCID: PMC4987931

PMID: 27141966

180. Nucleic Acids Res. 2016 Jul 8;44(W1):W536-41. doi: 10.1093/nar/gkw385. Epub 2016

May 3.

Peptiderive server: derive peptide inhibitors from protein-protein interactions.

Sedan Y(1), Marcu O(2), Lyskov S(3), Schueler-Furman O(4).

Author information:

(1)Department of Microbiology and Molecular Genetics, Faculty of Medicine, Hebrew

University, Jerusalem 91120, Israel Department of Chemical and Biomolecular

Engineering, John Hopkins University, Baltimore, MD 21218, USA. (2)Department of

Microbiology and Molecular Genetics, Faculty of Medicine, Hebrew University,

Jerusalem 91120, Israel. (3)Racah Institute of Physics, Hebrew University of

Jerusalem, Israel. (4)Department of Microbiology and Molecular Genetics, Faculty

of Medicine, Hebrew University, Jerusalem 91120, Israel oraf@ekmd.huji.ac.il.

The Rosetta Peptiderive protocol identifies, in a given structure of a

protein-protein interaction, the linear polypeptide segment suggested to

contribute most to binding energy. Interactions that feature a 'hot segment', a

linear peptide with significant binding energy compared to that of the complex,

may be amenable for inhibition and the peptide sequence and structure derived

from the interaction provide a starting point for rational drug design. Here we

present a web server for Peptiderive, which is incorporated within the ROSIE web

interface for Rosetta protocols. A new feature of the protocol also evaluates

whether derived peptides are good candidates for cyclization. Fast computation

times and clear visualization allow users to quickly assess the interaction of

interest. The Peptiderive server is available for free use at

http://rosie.rosettacommons.org/peptiderive.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw385

PMCID: PMC4987930

PMID: 27141963

181. Nucleic Acids Res. 2016 Jul 8;44(W1):W90-7. doi: 10.1093/nar/gkw377. Epub 2016

May 3.

Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.

Kuleshov MV(1), Jones MR(1), Rouillard AD(1), Fernandez NF(1), Duan Q(1), Wang

Z(1), Koplev S(1), Jenkins SL(1), Jagodnik KM(2), Lachmann A(1), McDermott MG(1),

Monteiro CD(1), Gundersen GW(1), Ma'ayan A(3).

Author information:

(1)Department of Pharmacology and Systems Therapeutics, BD2K-LINCS Data

Coordination and Integration Center, Icahn School of Medicine at Mount Sinai, One

Gustave L. Levy Place Box 1215, New York, NY 10029, USA. (2)Fluid Physics and

Transport Processes Branch, NASA Glenn Research Center, 21000 Brookpark Rd.,

Cleveland, OH 44135, USA. (3)Department of Pharmacology and Systems Therapeutics,

BD2K-LINCS Data Coordination and Integration Center, Icahn School of Medicine at

Mount Sinai, One Gustave L. Levy Place Box 1215, New York, NY 10029, USA

avi.maayan@mssm.edu.

Enrichment analysis is a popular method for analyzing gene sets generated by

genome-wide experiments. Here we present a significant update to one of the tools

in this domain called Enrichr. Enrichr currently contains a large collection of

diverse gene set libraries available for analysis and download. In total, Enrichr

currently contains 180 184 annotated gene sets from 102 gene set libraries. New

features have been added to Enrichr including the ability to submit fuzzy sets,

upload BED files, improved application programming interface and visualization of

the results as clustergrams. Overall, Enrichr is a comprehensive resource for

curated gene sets and a search engine that accumulates biological knowledge for

further biological discoveries. Enrichr is freely available at:

http://amp.pharm.mssm.edu/Enrichr.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw377

PMCID: PMC4987924

PMID: 27141961

182. Nucleic Acids Res. 2016 Jul 8;44(W1):W294-301. doi: 10.1093/nar/gkw362. Epub 2016

May 2.

RNAex: an RNA secondary structure prediction server enhanced by high-throughput

structure-probing data.

Wu Y(1), Qu R(1), Huang Y(1), Shi B(1), Liu M(1), Li Y(1), Lu ZJ(2).

Author information:

(1)MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems

Biology, Center for Plant Biology and Tsinghua-Peking Joint Center for Life

Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China.

(2)MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems

Biology, Center for Plant Biology and Tsinghua-Peking Joint Center for Life

Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China

zhilu@tsinghua.edu.cn.

Several high-throughput technologies have been developed to probe RNA base pairs

and loops at the transcriptome level in multiple species. However, to obtain the

final RNA secondary structure, extensive effort and considerable expertise is

required to statistically process the probing data and combine them with free

energy models. Therefore, we developed an RNA secondary structure prediction

server that is enhanced by experimental data (RNAex). RNAex is a web interface

that enables non-specialists to easily access cutting-edge structure-probing data

and predict RNA secondary structures enhanced by in vivo and in vitro data. RNAex

annotates the RNA editing, RNA modification and SNP sites on the predicted

structures. It provides four structure-folding methods, restrained MaxExpect,

SeqFold, RNAstructure (Fold) and RNAfold that can be selected by the user. The

performance of these four folding methods has been verified by previous

publications on known structures. We re-mapped the raw sequencing data of the

probing experiments to the whole genome for each species. RNAex thus enables

users to predict secondary structures for both known and novel RNA transcripts in

human, mouse, yeast and Arabidopsis The RNAex web server is available at

http://RNAex.ncrnalab.org/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw362

PMCID: PMC4987914

PMID: 27137891

183. Nucleic Acids Res. 2016 Jul 8;44(W1):W212-6. doi: 10.1093/nar/gkw369. Epub 2016

May 2.

Actionable pathways: interactive discovery of therapeutic targets using signaling

pathway models.

Salavert F(1), Hidago MR(2), Amadoz A(2), Çubuk C(2), Medina I(3), Crespo D(2),

Carbonell-Caballero J(2), Dopazo J(4).

Author information:

(1)Computational Genomics Department, Centro de Investigación Príncipe Felipe

(CIPF), Valencia, 46012, Spain Bioinformatics in Rare Diseases (BiER), Centro de

Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Valencia, 46012,

Spain. (2)Computational Genomics Department, Centro de Investigación Príncipe

Felipe (CIPF), Valencia, 46012, Spain. (3)HPC Service, University Information

Services, University of Cambridge, Cambridge, CB3 0RB, UK. (4)Computational

Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia,

46012, Spain Bioinformatics in Rare Diseases (BiER), Centro de Investigación

Biomédica en Red de Enfermedades Raras (CIBERER), Valencia, 46012, Spain

Functional Genomics Node, (INB, PRB2, ISCIII) at CIPF, Valencia 46012, Spain

jdopazo@cipf.es.

The discovery of actionable targets is crucial for targeted therapies and is also

a constituent part of the drug discovery process. The success of an intervention

over a target depends critically on its contribution, within the complex network

of gene interactions, to the cellular processes responsible for disease

progression or therapeutic response. Here we present PathAct, a web server that

predicts the effect that interventions over genes (inhibitions or activations

that simulate knock-outs, drug treatments or over-expressions) can have over

signal transmission within signaling pathways and, ultimately, over the cell

functionalities triggered by them. PathAct implements an advanced graphical

interface that provides a unique interactive working environment in which the

suitability of potentially actionable genes, that could eventually become drug

targets for personalized or individualized therapies, can be easily tested. The

PathAct tool can be found at: http://pathact.babelomics.org.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw369

PMCID: PMC4987920

PMID: 27137885

184. Nucleic Acids Res. 2016 Jul 8;44(W1):W529-35. doi: 10.1093/nar/gkw363. Epub 2016

Apr 30.

APID interactomes: providing proteome-based interactomes with controlled quality

for multiple species and derived networks.

Alonso-López D(1), Gutiérrez MA(1), Lopes KP(1), Prieto C(1), Santamaría R(1), De

Las Rivas J(2).

Author information:

(1)Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), Consejo Superior de

Investigaciones Científicas (CSIC) and Universidad de Salamanca (USAL), 37007

Salamanca, Spain. (2)Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), Consejo

Superior de Investigaciones Científicas (CSIC) and Universidad de Salamanca

(USAL), 37007 Salamanca, Spain jrivas@usal.es.

APID (Agile Protein Interactomes DataServer) is an interactive web server that

provides unified generation and delivery of protein interactomes mapped to their

respective proteomes. This resource is a new, fully redesigned server that

includes a comprehensive collection of protein interactomes for more than 400

organisms (25 of which include more than 500 interactions) produced by the

integration of only experimentally validated protein-protein physical

interactions. For each protein-protein interaction (PPI) the server includes

currently reported information about its experimental validation to allow

selection and filtering at different quality levels. As a whole, it provides easy

access to the interactomes from specific species and includes a global uniform

compendium of 90,379 distinct proteins and 678,441 singular interactions. APID

integrates and unifies PPIs from major primary databases of molecular

interactions, from other specific repositories and also from experimentally

resolved 3D structures of protein complexes where more than two proteins were

identified. For this purpose, a collection of 8,388 structures were analyzed to

identify specific PPIs. APID also includes a new graph tool (based on

Cytoscape.js) for visualization and interactive analyses of PPI networks. The

server does not require registration and it is freely available for use at

http://apid.dep.usal.es.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw363

PMCID: PMC4987915

PMID: 27131791

185. Nucleic Acids Res. 2016 Jul 8;44(W1):W522-8. doi: 10.1093/nar/gkw366. Epub 2016

Apr 30.

PepComposer: computational design of peptides binding to a given protein surface.

Obarska-Kosinska A(1), Iacoangeli A(1), Lepore R(2), Tramontano A(3).

Author information:

(1)Department of Physics, Sapienza University, Piazzale Aldo Moro, 5-00184 Rome,

Italy. (2)Department of Physics, Sapienza University, Piazzale Aldo Moro, 5-00184

Rome, Italy alba.lepore@uniroma1.it. (3)Department of Physics, Sapienza

University, Piazzale Aldo Moro, 5-00184 Rome, Italy Istituto Pasteur-Fondazione

Cenci Bolognetti, Viale Regina Elena 291, 00161 Rome, Italy.

There is a wide interest in designing peptides able to bind to a specific region

of a protein with the aim of interfering with a known interaction or as starting

point for the design of inhibitors. Here we describe PepComposer, a new pipeline

for the computational design of peptides binding to a given protein surface.

PepComposer only requires the target protein structure and an approximate

definition of the binding site as input. We first retrieve a set of peptide

backbone scaffolds from monomeric proteins that harbor the same backbone

arrangement as the binding site of the protein of interest. Next, we design

optimal sequences for the identified peptide scaffolds. The method is fully

automatic and available as a web server at

http://biocomputing.it/pepcomposer/webserver.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw366

PMCID: PMC4987918

PMID: 27131789

186. Nucleic Acids Res. 2016 Jul 8;44(W1):W154-9. doi: 10.1093/nar/gkw378. Epub 2016

Apr 30.

ICM: a web server for integrated clustering of multi-dimensional biomedical data.

He S(1), He H(1), Xu W(1), Huang X(1), Jiang S(1), Li F(2), He F(3), Bo X(4).

Author information:

(1)Beijing Institute of Radiation Medicine, Beijing 100850, P. R. China.

(2)Beijing Institute of Radiation Medicine, Beijing 100850, P. R. China

pittacus@gmail.com. (3)Beijing Institute of Radiation Medicine, Beijing 100850,

P. R. China hefc@bmi.ac.cn. (4)Beijing Institute of Radiation Medicine, Beijing

100850, P. R. China boxiaoc@163.com.

Large-scale efforts for parallel acquisition of multi-omics profiling continue to

generate extensive amounts of multi-dimensional biomedical data. Thus, integrated

clustering of multiple types of omics data is essential for developing

individual-based treatments and precision medicine. However, while rapid progress

has been made, methods for integrated clustering are lacking an intuitive web

interface that facilitates the biomedical researchers without sufficient

programming skills. Here, we present a web tool, named Integrated Clustering of

Multi-dimensional biomedical data (ICM), that provides an interface from which to

fuse, cluster and visualize multi-dimensional biomedical data and knowledge. With

ICM, users can explore the heterogeneity of a disease or a biological process by

identifying subgroups of patients. The results obtained can then be interactively

modified by using an intuitive user interface. Researchers can also exchange the

results from ICM with collaborators via a web link containing a Project ID number

that will directly pull up the analysis results being shared. ICM also support

incremental clustering that allows users to add new sample data into the data of

a previous study to obtain a clustering result. Currently, the ICM web server is

available with no login requirement and at no cost at

http://biotech.bmi.ac.cn/icm/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw378

PMCID: PMC4987925

PMID: 27131784

187. Nucleic Acids Res. 2016 Jul 8;44(W1):W507-13. doi: 10.1093/nar/gkw335. Epub 2016

Apr 29.

systemsDock: a web server for network pharmacology-based prediction and analysis.

Hsin KY(1), Matsuoka Y(2), Asai Y(3), Kamiyoshi K(3), Watanabe T(4), Kawaoka

Y(5), Kitano H(6).

Author information:

(1)Integrated Open Systems Unit, Okinawa Institute of Science and Technology

Graduate University, Onna-son, Okinawa 904-0412, Japan kunyi.hsin@oist.jp. (2)The

Systems Biology Institute, Minato, Tokyo 108-0071, Japan. (3)Integrated Open

Systems Unit, Okinawa Institute of Science and Technology Graduate University,

Onna-son, Okinawa 904-0412, Japan. (4)Division of Virology, Department of

Microbiology and Immunology, Institute of Medical Science, The University of

Tokyo, Minato, Tokyo 108-8639, Japan. (5)Division of Virology, Department of

Microbiology and Immunology, Institute of Medical Science, The University of

Tokyo, Minato, Tokyo 108-8639, Japan Department of Pathobiological Sciences,

Influenza Research Institute, School of Veterinary Medicine, University of

Wisconsin-Madison, Madison, Wisconsin 53711, USA. (6)Integrated Open Systems

Unit, Okinawa Institute of Science and Technology Graduate University, Onna-son,

Okinawa 904-0412, Japan The Systems Biology Institute, Minato, Tokyo 108-0071,

Japan Laboratory for Disease Systems Modeling, RIKEN Center for Integrative

Medical Sciences, Yokohama City, Kanagawa 230-0045, Japan kitano@sbi.jp.

We present systemsDock, a web server for network pharmacology-based prediction

and analysis, which permits docking simulation and molecular pathway map for

comprehensive characterization of ligand selectivity and interpretation of ligand

action on a complex molecular network. It incorporates an elaborately designed

scoring function for molecular docking to assess protein-ligand binding

potential. For large-scale screening and ease of investigation, systemsDock has a

user-friendly GUI interface for molecule preparation, parameter specification and

result inspection. Ligand binding potentials against individual proteins can be

directly displayed on an uploaded molecular interaction map, allowing users to

systemically investigate network-dependent effects of a drug or drug candidate. A

case study is given to demonstrate how systemsDock can be used to discover a test

compound's multi-target activity. systemsDock is freely accessible at

http://systemsdock.unit.oist.jp/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw335

PMCID: PMC4987901

PMID: 27131384

188. Nucleic Acids Res. 2016 Jul 8;44(W1):W383-9. doi: 10.1093/nar/gkw308. Epub 2016

Apr 29.

LassoProt: server to analyze biopolymers with lassos.

Dabrowski-Tumanski P(1), Niemyska W(2), Pasznik P(3), Sulkowska JI(4).

Author information:

(1)University of Warsaw, Faculty of Chemistry, Pasteura 1, Warsaw, Poland Centre

of New Technologies, University of Warsaw, Banacha 2c, Warsaw, Poland. (2)Centre

of New Technologies, University of Warsaw, Banacha 2c, Warsaw, Poland University

of Silesia, Institute of Mathematics, Bankowa 14, Katowice, Poland. (3)Centre of

New Technologies, University of Warsaw, Banacha 2c, Warsaw, Poland. (4)University

of Warsaw, Faculty of Chemistry, Pasteura 1, Warsaw, Poland Centre of New

Technologies, University of Warsaw, Banacha 2c, Warsaw, Poland

jsulkowska@chem.uw.edu.pl.

The LassoProt server, http://lassoprot.cent.uw.edu.pl/, enables analysis of

biopolymers with entangled configurations called lassos. The server offers

various ways of visualizing lasso configurations, as well as their time

trajectories, with all the results and plots downloadable. Broad spectrum of

applications makes LassoProt a useful tool for biologists, biophysicists,

chemists, polymer physicists and mathematicians. The server and our methods have

been validated on the whole PDB, and the results constitute the database of

proteins with complex lassos, supported with basic biological data. This database

can serve as a source of information about protein geometry and

entanglement-function correlations, as a reference set in protein modeling, and

for many other purposes.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw308

PMCID: PMC4987892

PMID: 27131383

189. Nucleic Acids Res. 2016 Jul 8;44(W1):W474-8. doi: 10.1093/nar/gkw361. Epub 2016

Apr 29.

SAbPred: a structure-based antibody prediction server.

Dunbar J(1), Krawczyk K(2), Leem J(2), Marks C(2), Nowak J(2), Regep C(2),

Georges G(3), Kelm S(4), Popovic B(5), Deane CM(6).

Author information:

(1)Department of Statistics, Oxford Protein Informatics Group, University of

Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK Pharma Research and Early

Development, Large Molecule Research, Roche Innovation Center Munich, Nonnenwald

2, 82377, Penzberg, Germany. (2)Department of Statistics, Oxford Protein

Informatics Group, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK.

(3)Pharma Research and Early Development, Large Molecule Research, Roche

Innovation Center Munich, Nonnenwald 2, 82377, Penzberg, Germany. (4)Informatics

Department, UCB Pharma, 208 Bath Road, Slough, SL1 3WE, UK. (5)Antibody Discovery

and Protein Engineering, Medimmune Ltd, Granta Park, Cambridge, CB21 6GH, UK.

(6)Department of Statistics, Oxford Protein Informatics Group, University of

Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK deane@stats.ox.ac.uk.

SAbPred is a server that makes predictions of the properties of antibodies

focusing on their structures. Antibody informatics tools can help improve our

understanding of immune responses to disease and aid in the design and

engineering of therapeutic molecules. SAbPred is a single platform containing

multiple applications which can: number and align sequences; automatically

generate antibody variable fragment homology models; annotate such models with

estimated accuracy alongside sequence and structural properties including

potential developability issues; predict paratope residues; and predict epitope

patches on protein antigens. The server is available at

http://opig.stats.ox.ac.uk/webapps/sabpred.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw361

PMCID: PMC4987913

PMID: 27131379

190. Nucleic Acids Res. 2016 Jul 8;44(W1):W351-5. doi: 10.1093/nar/gkw357. Epub 2016

Apr 29.

Dali server update.

Holm L(1), Laakso LM(2).

Author information:

(1)Institute of Biotechnology, University of Helsinki, PO Box 56, Finland

Department of Biosciences, University of Helsinki, PO Box 56, Finland

Liisa.Holm@helsinki.fi. (2)Institute of Biotechnology, University of Helsinki, PO

Box 56, Finland.

The Dali server (http://ekhidna2.biocenter.helsinki.fi/dali) is a network service

for comparing protein structures in 3D. In favourable cases, comparing 3D

structures may reveal biologically interesting similarities that are not

detectable by comparing sequences. The Dali server has been running in various

places for over 20 years and is used routinely by crystallographers on newly

solved structures. The latest update of the server provides enhanced analytics

for the study of sequence and structure conservation. The server performs three

types of structure comparisons: (i) Protein Data Bank (PDB) search compares one

query structure against those in the PDB and returns a list of similar

structures; (ii) pairwise comparison compares one query structure against a list

of structures specified by the user; and (iii) all against all structure

comparison returns a structural similarity matrix, a dendrogram and a

multidimensional scaling projection of a set of structures specified by the user.

Structural superimpositions are visualized using the Java-free WebGL viewer PV.

The structural alignment view is enhanced by sequence similarity searches against

Uniprot. The combined structure-sequence alignment information is compressed to a

stack of aligned sequence logos. In the stack, each structure is structurally

aligned to the query protein and represented by a sequence logo.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw357

PMCID: PMC4987910

PMID: 27131377

191. Nucleic Acids Res. 2016 Jul 8;44(W1):W217-25. doi: 10.1093/nar/gkw342. Epub 2016

Apr 29.

MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design

with competing endogenous reactions in mind.

Kuwahara H(1), Alazmi M(1), Cui X(1), Gao X(2).

Author information:

(1)King Abdullah University of Science and Technology (KAUST), Computational

Bioscience Research Center (CBRC), Thuwal, 23955, Saudi Arabia. (2)King Abdullah

University of Science and Technology (KAUST), Computational Bioscience Research

Center (CBRC), Thuwal, 23955, Saudi Arabia xin.gao@kaust.edu.sa.

To rationally design a productive heterologous biosynthesis system, it is

essential to consider the suitability of foreign reactions for the specific

endogenous metabolic infrastructure of a host. We developed a novel web server,

called MRE, which, for a given pair of starting and desired compounds in a given

chassis organism, ranks biosynthesis routes from the perspective of the

integration of new reactions into the endogenous metabolic system. For each

promising heterologous biosynthesis pathway, MRE suggests actual enzymes for

foreign metabolic reactions and generates information on competing endogenous

reactions for the consumption of metabolites. These unique, chassis-centered

features distinguish MRE from existing pathway design tools and allow synthetic

biologists to evaluate the design of their biosynthesis systems from a different

angle. By using biosynthesis of a range of high-value natural products as a case

study, we show that MRE is an effective tool to guide the design and optimization

of heterologous biosynthesis pathways. The URL of MRE is

http://www.cbrc.kaust.edu.sa/mre/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw342

PMCID: PMC4987905

PMID: 27131375

192. Nucleic Acids Res. 2016 Jul 8;44(W1):W406-9. doi: 10.1093/nar/gkw336. Epub 2016

Apr 29.

3Drefine: an interactive web server for efficient protein structure refinement.

Bhattacharya D(1), Nowotny J(1), Cao R(1), Cheng J(2).

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, MO 65211,

USA. (2)Department of Computer Science, University of Missouri, Columbia, MO

65211, USA Informatics Institute, University of Missouri, Columbia, MO 65211, USA

C. Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA

chengji@missouri.edu.

3Drefine is an interactive web server for consistent and computationally

efficient protein structure refinement with the capability to perform web-based

statistical and visual analysis. The 3Drefine refinement protocol utilizes

iterative optimization of hydrogen bonding network combined with atomic-level

energy minimization on the optimized model using a composite physics and

knowledge-based force fields for efficient protein structure refinement. The

method has been extensively evaluated on blind CASP experiments as well as on

large-scale and diverse benchmark datasets and exhibits consistent improvement

over the initial structure in both global and local structural quality measures.

The 3Drefine web server allows for convenient protein structure refinement

through a text or file input submission, email notification, provided example

submission and is freely available without any registration requirement. The

server also provides comprehensive analysis of submissions through various energy

and statistical feedback and interactive visualization of multiple refined models

through the JSmol applet that is equipped with numerous protein model analysis

tools. The web server has been extensively tested and used by many users. As a

result, the 3Drefine web server conveniently provides a useful tool easily

accessible to the community. The 3Drefine web server has been made publicly

available at the URL: http://sysbio.rnet.missouri.edu/3Drefine/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw336

PMCID: PMC4987902

PMID: 27131371

193. Nucleic Acids Res. 2016 Jul 8;44(W1):W542-9. doi: 10.1093/nar/gkw340. Epub 2016

Apr 29.

InterEvDock: a docking server to predict the structure of protein-protein

interactions using evolutionary information.

Yu J(1), Vavrusa M(2), Andreani J(1), Rey J(2), Tufféry P(3), Guerois R(4).

Author information:

(1)Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ

Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette cedex, France. (2)INSERM

UMR-S 973 Molécules Thérapeutiques in Silico, INSERM UMR-S 973, RPBS, Université

Paris Diderot, 35 rue H. Brion, case 7113, Sorbone Paris Cité, 75205 Paris cedex

13, France. (3)INSERM UMR-S 973 Molécules Thérapeutiques in Silico, INSERM UMR-S

973, RPBS, Université Paris Diderot, 35 rue H. Brion, case 7113, Sorbone Paris

Cité, 75205 Paris cedex 13, France pierre.tuffery@univ-paris-diderot.fr.

(4)Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ

Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette cedex, France

guerois@cea.fr.

The structural modeling of protein-protein interactions is key in understanding

how cell machineries cross-talk with each other. Molecular docking simulations

provide efficient means to explore how two unbound protein structures interact.

InterEvDock is a server for protein docking based on a free rigid-body docking

strategy. A systematic rigid-body docking search is performed using the FRODOCK

program and the resulting models are re-scored with InterEvScore and SOAP-PP

statistical potentials. The InterEvScore potential was specifically designed to

integrate co-evolutionary information in the docking process. InterEvDock server

is thus particularly well suited in case homologous sequences are available for

both binding partners. The server returns 10 structures of the most likely

consensus models together with 10 predicted residues most likely involved in the

interface. In 91% of all complexes tested in the benchmark, at least one residue

out of the 10 predicted is involved in the interface, providing useful guidelines

for mutagenesis. InterEvDock is able to identify a correct model among the top10

models for 49% of the rigid-body cases with evolutionary information, making it a

unique and efficient tool to explore structural interactomes under an

evolutionary perspective. The InterEvDock web interface is available at

http://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw340

PMCID: PMC4987904

PMID: 27131368

194. Nucleic Acids Res. 2016 Jul 8;44(W1):W502-6. doi: 10.1093/nar/gkw360. Epub 2016

Apr 29.

Galaxy7TM: flexible GPCR-ligand docking by structure refinement.

Lee GR(1), Seok C(2).

Author information:

(1)Department of Chemistry, Seoul National University, Seoul 151-747, Korea.

(2)Department of Chemistry, Seoul National University, Seoul 151-747, Korea

chaok@snu.ac.kr.

G-protein-coupled receptors (GPCRs) play important physiological roles related to

signal transduction and form a major group of drug targets. Prediction of

GPCR-ligand complex structures has therefore important implications to drug

discovery. With previously available servers, it was only possible to first

predict GPCR structures by homology modeling and then perform ligand docking on

the model structures. However, model structures generated without explicit

consideration of specific ligands of interest can be inaccurate because GPCR

structures can be affected by ligand binding. The Galaxy7TM server, freely

accessible at http://galaxy.seoklab.org/7TM, improves an input GPCR structure by

simultaneous ligand docking and flexible structure refinement using GALAXY

methods. The server shows better performance in both ligand docking and GPCR

structure refinement than commonly used programs AutoDock Vina and Rosetta

MPrelax, respectively.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw360

PMCID: PMC4987912

PMID: 27131365

195. Nucleic Acids Res. 2016 Jul 8;44(W1):W514-21. doi: 10.1093/nar/gkw314. Epub 2016

Apr 29.

MANORAA (Mapping Analogous Nuclei Onto Residue And Affinity) for identifying

protein-ligand fragment interaction, pathways and SNPs.

Tanramluk D(1), Narupiyakul L(2), Akavipat R(3), Gong S(4), Charoensawan V(5).

Author information:

(1)Institute of Molecular Biosciences, Mahidol University, Salaya, Nakhon Pathom

73170, Thailand Integrative Computational BioScience (ICBS) Center, Mahidol

University, Salaya, Nakhon Pathom 73170, Thailand duangrudee.tan@mahidol.ac.th.

(2)Integrative Computational BioScience (ICBS) Center, Mahidol University,

Salaya, Nakhon Pathom 73170, Thailand Department of Computer Engineering, Faculty

of Engineering, Mahidol University, Salaya, Nakhon Pathom 73170, Thailand.

(3)Integrative Computational BioScience (ICBS) Center, Mahidol University,

Salaya, Nakhon Pathom 73170, Thailand Department of Computer Science, Faculty of

Science, Kasetsart University, Chatuchak, Bangkok 10900, Thailand. (4)Department

of Obstetrics and Gynaecology, University of Cambridge, The Rosie Hospital,

Cambridge CB2 0SW, UK. (5)Integrative Computational BioScience (ICBS) Center,

Mahidol University, Salaya, Nakhon Pathom 73170, Thailand Department of

Biochemistry, Faculty of Science, Mahidol University, Ratchathewi, Bangkok 10400,

Thailand.

Protein-ligand interaction analysis is an important step of drug design and

protein engineering in order to predict the binding affinity and selectivity

between ligands to the target proteins. To date, there are more than 100 000

structures available in the Protein Data Bank (PDB), of which ∼30% are

protein-ligand (MW below 1000 Da) complexes. We have developed the integrative

web server MANORAA (Mapping Analogous Nuclei Onto Residue And Affinity) with the

aim of providing a user-friendly web interface to assist structural study and

design of protein-ligand interactions. In brief, the server allows the users to

input the chemical fragments and present all the unique molecular interactions to

the target proteins with available three-dimensional structures in the PDB. The

users can also link the ligands of interest to assess possible off-target

proteins, human variants and pathway information using our all-in-one integrated

tools. Taken together, we envisage that the server will facilitate and improve

the study of protein-ligand interactions by allowing observation and comparison

of ligand interactions with multiple proteins at the same time.

(http://manoraa.org).

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw314

PMCID: PMC4987895

PMID: 27131358

196. Nucleic Acids Res. 2016 Jul 8;44(W1):W302-7. doi: 10.1093/nar/gkw337. Epub 2016

Apr 29.

Rtools: a web server for various secondary structural analyses on single RNA

sequences.

Hamada M(1), Ono Y(2), Kiryu H(3), Sato K(4), Kato Y(5), Fukunaga T(3), Mori

R(3), Asai K(6).

Author information:

(1)Department of Electrical Engineering and Bioscience, Faculty of Science and

Engineering, Waseda University, 55N-06-10, 3-4-1, Okubo Shinjuku-ku, Tokyo

169-8555, Japan Artificial Intelligence Research Center (AIRC), National

Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi,

Koto-ku, 135-0064 Tokyo, Japan. (2)IMSBIO Co., Ltd, 4-21-1-601 Higashi-Ikebukuro,

Toshima-ku, Tokyo 170-0013, Japan. (3)Graduate School of Frontier Sciences,

University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-8562, Japan. (4)Department of

Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama

223-8522, Japan. (5)Center for iPS Cell Research and Application (CiRA), Kyoto

University, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan. (6)Artificial Intelligence

Research Center (AIRC), National Institute of Advanced Industrial Science and

Technology (AIST), 2-4-7 Aomi, Koto-ku, 135-0064 Tokyo, Japan Graduate School of

Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-8562,

Japan asai@k.u-tokyo.ac.jp.

The secondary structures, as well as the nucleotide sequences, are the important

features of RNA molecules to characterize their functions. According to the

thermodynamic model, however, the probability of any secondary structure is very

small. As a consequence, any tool to predict the secondary structures of RNAs has

limited accuracy. On the other hand, there are a few tools to compensate the

imperfect predictions by calculating and visualizing the secondary structural

information from RNA sequences. It is desirable to obtain the rich information

from those tools through a friendly interface. We implemented a web server of the

tools to predict secondary structures and to calculate various structural

features based on the energy models of secondary structures. By just giving an

RNA sequence to the web server, the user can get the different types of solutions

of the secondary structures, the marginal probabilities such as base-paring

probabilities, loop probabilities and accessibilities of the local bases, the

energy changes by arbitrary base mutations as well as the measures for

validations of the predicted secondary structures. The web server is available at

http://rtools.cbrc.jp, which integrates software tools, CentroidFold,

CentroidHomfold, IPKnot, CapR, Raccess, Rchange and RintD.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw337

PMCID: PMC4987903

PMID: 27131356

197. Nucleic Acids Res. 2016 Jul 8;44(W1):W430-5. doi: 10.1093/nar/gkw306. Epub 2016

Apr 25.

RaptorX-Property: a web server for protein structure property prediction.

Wang S(1), Li W(2), Liu S(2), Xu J(3).

Author information:

(1)Toyota Technological Institute at Chicago, Chicago, IL, USA Department of

Human Genetics, University of Chicago, Chicago, IL, USA wangsheng@uchicago.edu.

(2)School of Biological and Chemical Engineering, Zhejiang University of Science

and Technology, Zhejiang, China. (3)Toyota Technological Institute at Chicago,

Chicago, IL, USA jinboxu@gmail.com.

RaptorX Property (http://raptorx2.uchicago.edu/StructurePropertyPred/predict/) is

a web server predicting structure property of a protein sequence without using

any templates. It outperforms other servers, especially for proteins without

close homologs in PDB or with very sparse sequence profile (i.e. carries little

evolutionary information). This server employs a powerful in-house deep learning

model DeepCNF (Deep Convolutional Neural Fields) to predict secondary structure

(SS), solvent accessibility (ACC) and disorder regions (DISO). DeepCNF not only

models complex sequence-structure relationship by a deep hierarchical

architecture, but also interdependency between adjacent property labels. Our

experimental results show that, tested on CASP10, CASP11 and the other

benchmarks, this server can obtain ∼84% Q3 accuracy for 3-state SS, ∼72% Q8

accuracy for 8-state SS, ∼66% Q3 accuracy for 3-state solvent accessibility, and

∼0.89 area under the ROC curve (AUC) for disorder prediction.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw306

PMCID: PMC4987890

PMID: 27112573

198. Nucleic Acids Res. 2016 Jul 8;44(W1):W361-6. doi: 10.1093/nar/gkw307. Epub 2016

Apr 25.

CoinFold: a web server for protein contact prediction and contact-assisted

protein folding.

Wang S(1), Li W(2), Zhang R(3), Liu S(2), Xu J(4).

Author information:

(1)Toyota Technological Institute at Chicago, Chicago, IL, USA Department of

Human Genetics, University of Chicago, Chicago, IL, USA wangsheng@uchicago.edu.

(2)School of Biological and Chemical Engineering, Zhejiang University of Science

and Technology, Zhejiang, China. (3)Toyota Technological Institute at Chicago,

Chicago, IL, USA. (4)Toyota Technological Institute at Chicago, Chicago, IL, USA

jinboxu@gmail.com.

CoinFold (http://raptorx2.uchicago.edu/ContactMap/) is a web server for protein

contact prediction and contact-assisted de novo structure prediction. CoinFold

predicts contacts by integrating joint multi-family evolutionary coupling (EC)

analysis and supervised machine learning. This joint EC analysis is unique in

that it not only uses residue coevolution information in the target protein

family, but also that in the related families which may have divergent sequences

but similar folds. The supervised learning further improves contact prediction

accuracy by making use of sequence profile, contact (distance) potential and

other information. Finally, this server predicts tertiary structure of a sequence

by feeding its predicted contacts and secondary structure to the CNS suite.

Tested on the CASP and CAMEO targets, this server shows significant advantages

over existing ones of similar category in both contact and tertiary structure

prediction.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw307

PMCID: PMC4987891

PMID: 27112569

199. Nucleic Acids Res. 2016 Jul 8;44(W1):W226-31. doi: 10.1093/nar/gkw305. Epub 2016

Apr 22.

SensiPath: computer-aided design of sensing-enabling metabolic pathways.

Delépine B(1), Libis V(1), Carbonell P(2), Faulon JL(3).

Author information:

(1)iSSB, Genopole, CNRS, UEVE, Université Paris Saclay, 91000 Évry, France

Micalis Institute, INRA, AgroParisTech, Université Paris Saclay, 78350

Jouy-en-Josas, France. (2)SYNBIOCHEM Centre, Manchester Institute of

Biotechnology, University of Manchester, M1 7DN Manchester, UK. (3)iSSB,

Genopole, CNRS, UEVE, Université Paris Saclay, 91000 Évry, France Micalis

Institute, INRA, AgroParisTech, Université Paris Saclay, 78350 Jouy-en-Josas,

France SYNBIOCHEM Centre, Manchester Institute of Biotechnology, University of

Manchester, M1 7DN Manchester, UK jlfaulon@jouy.inra.fr.

Genetically-encoded biosensors offer a wide range of opportunities to develop

advanced synthetic biology applications. Circuits with the ability of detecting

and quantifying intracellular amounts of a compound of interest are central to

whole-cell biosensors design for medical and environmental applications, and they

also constitute essential parts for the selection and regulation of high-producer

strains in metabolic engineering. However, the number of compounds that can be

detected through natural mechanisms, like allosteric transcription factors, is

limited; expanding the set of detectable compounds is therefore highly desirable.

Here, we present the SensiPath web server, accessible at

http://sensipath.micalis.fr SensiPath implements a strategy to enlarge the set of

detectable compounds by screening for multi-step enzymatic transformations

converting non-detectable compounds into detectable ones. The SensiPath approach

is based on the encoding of reactions through signature descriptors to explore

sensing-enabling metabolic pathways, which are putative biochemical

transformations of the target compound leading to known effectors of

transcription factors. In that way, SensiPath enlarges the design space by

broadening the potential use of biosensors in synthetic biology applications.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw305

PMID: 27106061

200. Nucleic Acids Res. 2016 Jul 8;44(W1):W339-43. doi: 10.1093/nar/gkw300. Epub 2016

Apr 22.

PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of

regular and transmembrane proteins using homology extension on reduced databases.

Floden EW(1), Tommaso PD(1), Chatzou M(1), Magis C(1), Notredame C(1), Chang

JM(2).

Author information:

(1)Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and

Technology, Dr Aiguader 88, Barcelona 08003, Spain Universitat Pompeu Fabra

(UPF), Barcelona 08003, Spain. (2)Department of Computer Science, National

Chengchi University, Taipei 11605, Taiwan chang.jiaming@gmail.com.

The PSI/TM-Coffee web server performs multiple sequence alignment (MSA) of

proteins by combining homology extension with a consistency based alignment

approach. Homology extension is performed with Position Specific Iterative (PSI)

BLAST searches against a choice of redundant and non-redundant databases. The

main novelty of this server is to allow databases of reduced complexity to

rapidly perform homology extension. This server also gives the possibility to use

transmembrane proteins (TMPs) reference databases to allow even faster homology

extension on this important category of proteins. Aside from an MSA, the server

also outputs topological prediction of TMPs using the HMMTOP algorithm. Previous

benchmarking of the method has shown this approach outperforms the most accurate

alignment methods such as MSAProbs, Kalign, PROMALS, MAFFT, ProbCons and

PRALINE™. The web server is available at http://tcoffee.crg.cat/tmcoffee.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw300

PMCID: PMC4987888

PMID: 27106060

201. Nucleic Acids Res. 2016 Jul 8;44(W1):W436-41. doi: 10.1093/nar/gkw320. Epub 2016

Apr 22.

USR-VS: a web server for large-scale prospective virtual screening using

ultrafast shape recognition techniques.

Li H(1), Leung KS(2), Wong MH(2), Ballester PJ(3).

Author information:

(1)Institute of Future Cities, Chinese University of Hong Kong, Hong Kong.

(2)Department of Computer Science and Engineering, Chinese University of Hong

Kong, Sha Tin, New Territories, Hong Kong. (3)Cancer Research Center of

Marseille, INSERM U1068, 13009-Marseille, France pedro.ballester@inserm.fr.

Ligand-based Virtual Screening (VS) methods aim at identifying molecules with a

similar activity profile across phenotypic and macromolecular targets to that of

a query molecule used as search template. VS using 3D similarity methods have the

advantage of biasing this search toward active molecules with innovative chemical

scaffolds, which are highly sought after in drug design to provide novel leads

with improved properties over the query molecule (e.g. patentable, of lower

toxicity or increased potency). Ultrafast Shape Recognition (USR) has

demonstrated excellent performance in the discovery of molecules with

previously-unknown phenotypic or target activity, with retrospective studies

suggesting that its pharmacophoric extension (USRCAT) should obtain even better

hit rates once it is used prospectively. Here we present USR-VS

(http://usr.marseille.inserm.fr/), the first web server using these two validated

ligand-based 3D methods for large-scale prospective VS. In about 2 s, 93.9

million 3D conformers, expanded from 23.1 million purchasable molecules, are

screened and the 100 most similar molecules among them in terms of 3D shape and

pharmacophoric properties are shown. USR-VS functionality also provides

interactive visualization of the similarity of the query molecule against the hit

molecules as well as vendor information to purchase selected hits in order to be

experimentally tested.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw320

PMCID: PMC4987897

PMID: 27106057

202. Nucleic Acids Res. 2016 Jul 8;44(W1):W70-6. doi: 10.1093/nar/gkw313. Epub 2016

Apr 22.

Genonets server-a web server for the construction, analysis and visualization of

genotype networks.

Khalid F(1), Aguilar-Rodríguez J(1), Wagner A(2), Payne JL(3).

Author information:

(1)Department of Evolutionary Biology and Environmental Studies, University of

Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland Swiss Institute of

Bioinformatics, Quartier Sorge-Batiment Genopode, 1015 Lausanne, Switzerland.

(2)Department of Evolutionary Biology and Environmental Studies, University of

Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland Swiss Institute of

Bioinformatics, Quartier Sorge-Batiment Genopode, 1015 Lausanne, Switzerland The

Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

andreas.wagner@ieu.uzh.ch. (3)Department of Evolutionary Biology and

Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057

Zurich, Switzerland Swiss Institute of Bioinformatics, Quartier Sorge-Batiment

Genopode, 1015 Lausanne, Switzerland joshua.payne@ieu.uzh.ch.

A genotype network is a graph in which vertices represent genotypes that have the

same phenotype. Edges connect vertices if their corresponding genotypes differ in

a single small mutation. Genotype networks are used to study the organization of

genotype spaces. They have shed light on the relationship between robustness and

evolvability in biological systems as different as RNA macromolecules and

transcriptional regulatory circuits. Despite the importance of genotype networks,

no tool exists for their automatic construction, analysis and visualization. Here

we fill this gap by presenting the Genonets Server, a tool that provides the

following features: (i) the construction of genotype networks for categorical and

univariate phenotypes from DNA, RNA, amino acid or binary sequences; (ii)

analyses of genotype network topology and how it relates to robustness and

evolvability, as well as analyses of genotype network topography and how it

relates to the navigability of a genotype network via mutation and natural

selection; (iii) multiple interactive visualizations that facilitate exploratory

research and education. The Genonets Server is freely available at

http://ieu-genonets.uzh.ch.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw313

PMCID: PMC4987894

PMID: 27106055

203. Nucleic Acids Res. 2016 Jul 8;44(W1):W390-4. doi: 10.1093/nar/gkw297. Epub 2016

Apr 21.

SL2: an interactive webtool for modeling of missing segments in proteins.

Ismer J(1), Rose AS(1), Tiemann JK(1), Goede A(2), Preissner R(2), Hildebrand

PW(3).

Author information:

(1)Institute of Medical Physics and Biophysics, University Medicine, Berlin,

10117 Berlin, Germany. (2)Institute of Physiology & Experimental Clinical

Research Center, University Medicine, Berlin, 13125, Germany. (3)Institute of

Medical Physics and Biophysics, University Medicine, Berlin, 10117 Berlin,

Germany peter.hildebrand@charite.de.

SuperLooper2 (SL2) (http://proteinformatics.charite.de/sl2) is the updated

version of our previous web-server SuperLooper, a fragment based tool for the

prediction and interactive placement of loop structures into globular and helical

membrane proteins. In comparison to our previous version, SL2 benefits from both

a considerably enlarged database of fragments derived from high-resolution 3D

protein structures of globular and helical membrane proteins, and the integration

of a new protein viewer. The database, now with double the content, significantly

improved the coverage of fragment conformations and prediction quality. The

employment of the NGL viewer for visualization of the protein under investigation

and interactive selection of appropriate loops makes SL2 independent of

third-party plug-ins and additional installations.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw297

PMCID: PMC4987885

PMID: 27105847

204. Nucleic Acids Res. 2016 Jul 8;44(W1):W29-34. doi: 10.1093/nar/gkw292. Epub 2016

Apr 21.

Companion: a web server for annotation and analysis of parasite genomes.

Steinbiss S(1), Silva-Franco F(2), Brunk B(3), Foth B(4), Hertz-Fowler C(2),

Berriman M(4), Otto TD(4).

Author information:

(1)Parasite Genomics, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10

1SA, UK ss34@sanger.ac.uk. (2)Institute of Integrative Biology, University of

Liverpool, Liverpool L69 7ZB, UK. (3)Department of Biology, University of

Pennsylvania, Philadelphia, PA 19104, USA. (4)Parasite Genomics, Wellcome Trust

Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.

Currently available sequencing technologies enable quick and economical

sequencing of many new eukaryotic parasite (apicomplexan or kinetoplastid)

species or strains. Compared to SNP calling approaches, de novo assembly of these

genomes enables researchers to additionally determine insertion, deletion and

recombination events as well as to detect complex sequence diversity, such as

that seen in variable multigene families. However, there currently are no

automated eukaryotic annotation pipelines offering the required range of results

to facilitate such analyses. A suitable pipeline needs to perform

evidence-supported gene finding as well as functional annotation and pseudogene

detection up to the generation of output ready to be submitted to a public

database. Moreover, no current tool includes quick yet informative comparative

analyses and a first pass visualization of both annotation and analysis results.

To overcome those needs we have developed the Companion web server

(http://companion.sanger.ac.uk) providing parasite genome annotation as a service

using a reference-based approach. We demonstrate the use and performance of

Companion by annotating two Leishmania and Plasmodium genomes as typical parasite

cases and evaluate the results compared to manually annotated references.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw292

PMCID: PMC4987884

PMID: 27105845

205. Nucleic Acids Res. 2016 Jul 8;44(W1):W142-6. doi: 10.1093/nar/gkw298. Epub 2016

Apr 21.

EXPLoRA-web: linkage analysis of quantitative trait loci using bulk segregant

analysis.

Pulido-Tamayo S(1), Duitama J(2), Marchal K(3).

Author information:

(1)Department of Information Technology, iGent Toren, Technologiepark 15, 9052

Gent, Belgium Department of Plant Biotechnology and Bioinformatics, UGent,

Technologiepark 927, 9052 Gent, Belgium Bioinformatics Institute Ghent,

Technologiepark 927, 9052 Gent, Belgium Department of Microbial and Molecular

Systems, KU Leuven, Kasteelpark Arenberg 20, B-3001 Leuven, Belgium.

(2)Agrobiodiversity Research Area, International Center for Tropical Agriculture

(CIAT), 763537 Cali, Colombia. (3)Department of Information Technology, iGent

Toren, Technologiepark 15, 9052 Gent, Belgium Department of Plant Biotechnology

and Bioinformatics, UGent, Technologiepark 927, 9052 Gent, Belgium Bioinformatics

Institute Ghent, Technologiepark 927, 9052 Gent, Belgium Department of Genetics,

University of Pretoria, Hatfield Campus, Pretoria 0028, South Africa

kathleen.marchal@intec.ugent.be.

Identification of genomic regions associated with a phenotype of interest is a

fundamental step toward solving questions in biology and improving industrial

research. Bulk segregant analysis (BSA) combined with high-throughput sequencing

is a technique to efficiently identify these genomic regions associated with a

trait of interest. However, distinguishing true from spuriously linked genomic

regions and accurately delineating the genomic positions of these truly linked

regions requires the use of complex statistical models currently implemented in

software tools that are generally difficult to operate for non-expert users. To

facilitate the exploration and analysis of data generated by bulked segregant

analysis, we present EXPLoRA-web, a web service wrapped around our previously

published algorithm EXPLoRA, which exploits linkage disequilibrium to increase

the power and accuracy of quantitative trait loci identification in BSA analysis.

EXPLoRA-web provides a user friendly interface that enables easy data upload and

parallel processing of different parameter configurations. Results are provided

graphically and as BED file and/or text file and the input is expected in widely

used formats, enabling straightforward BSA data analysis. The web server is

available at http://bioinformatics.intec.ugent.be/explora-web/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw298

PMCID: PMC4987886

PMID: 27105844

206. Nucleic Acids Res. 2016 Jul 8;44(W1):W22-8. doi: 10.1093/nar/gkw255. Epub 2016

Apr 20.

EDGAR 2.0: an enhanced software platform for comparative gene content analyses.

Blom J(1), Kreis J(2), Spänig S(2), Juhre T(2), Bertelli C(3), Ernst C(4),

Goesmann A(2).

Author information:

(1)Bioinformatics & Systems Biology, Justus-Liebig-University Giessen, 35392

Giessen, Hesse, Germany jochen.blom@computational.bio.uni-giessen.de.

(2)Bioinformatics & Systems Biology, Justus-Liebig-University Giessen, 35392

Giessen, Hesse, Germany. (3)Institute of Microbiology, University Hospital Center

and University of Lausanne, 1011 Lausanne, VD, Switzerland SIB Swiss Institute of

Bioinformatics, 1015 Lausanne, VD, Switzerland. (4)Center for Familial Breast and

Ovarian Cancer, Medical Faculty, University Hospital Cologne, University of

Cologne, 50931 Cologne, NRW, Germany.

The rapidly increasing availability of microbial genome sequences has led to a

growing demand for bioinformatics software tools that support the functional

analysis based on the comparison of closely related genomes. By utilizing

comparative approaches on gene level it is possible to gain insights into the

core genes which represent the set of shared features for a set of organisms

under study. Vice versa singleton genes can be identified to elucidate the

specific properties of an individual genome. Since initial publication, the EDGAR

platform has become one of the most established software tools in the field of

comparative genomics. Over the last years, the software has been continuously

improved and a large number of new analysis features have been added. For the new

version, EDGAR 2.0, the gene orthology estimation approach was newly designed and

completely re-implemented. Among other new features, EDGAR 2.0 provides extended

phylogenetic analysis features like AAI (Average Amino Acid Identity) and ANI

(Average Nucleotide Identity) matrices, genome set size statistics and modernized

visualizations like interactive synteny plots or Venn diagrams. Thereby, the

software supports a quick and user-friendly survey of evolutionary relationships

between microbial genomes and simplifies the process of obtaining new biological

insights into their differential gene content. All features are offered to the

scientific community via a web-based and therefore platform-independent user

interface, which allows easy browsing of precomputed datasets. The web server is

accessible at http://edgar.computational.bio.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw255

PMCID: PMC4987874

PMID: 27098043

207. Nucleic Acids Res. 2016 Jul 8;44(W1):W83-9. doi: 10.1093/nar/gkw199. Epub 2016

Apr 20.

g:Profiler-a web server for functional interpretation of gene lists (2016

update).

Reimand J(1), Arak T(2), Adler P(2), Kolberg L(2), Reisberg S(2), Peterson H(2),

Vilo J(3).

Author information:

(1)Ontario Institute for Cancer Research, 661 University Avenue, Toronto, ON M5G

0A3, Canada Department of Medical Biophysics, University of Toronto, 101 College

Street, Toronto, ON M5G 1L7, Canada. (2)Institute of Computer Science, University

of Tartu, Liivi 2, 50409 Tartu, Estonia. (3)Institute of Computer Science,

University of Tartu, Liivi 2, 50409 Tartu, Estonia vilo@ut.ee.

Functional enrichment analysis is a key step in interpreting gene lists

discovered in diverse high-throughput experiments. g:Profiler studies flat and

ranked gene lists and finds statistically significant Gene Ontology terms,

pathways and other gene function related terms. Translation of hundreds of gene

identifiers is another core feature of g:Profiler. Since its first publication in

2007, our web server has become a popular tool of choice among basic and

translational researchers. Timeliness is a major advantage of g:Profiler as

genome and pathway information is synchronized with the Ensembl database in

quarterly updates. g:Profiler supports 213 species including mammals and other

vertebrates, plants, insects and fungi. The 2016 update of g:Profiler introduces

several novel features. We have added further functional datasets to interpret

gene lists, including transcription factor binding site predictions, Mendelian

disease annotations, information about protein expression and complexes and gene

mappings of human genetic polymorphisms. Besides the interactive web interface,

g:Profiler can be accessed in computational pipelines using our R package, Python

interface and BioJS component. g:Profiler is freely available at

http://biit.cs.ut.ee/gprofiler/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw199

PMCID: PMC4987867

PMID: 27098042

208. Nucleic Acids Res. 2016 Jul 8;44(W1):W315-9. doi: 10.1093/nar/gkw279. Epub 2016

Apr 19.

SimRNAweb: a web server for RNA 3D structure modeling with optional restraints.

Magnus M(1), Boniecki MJ(1), Dawson W(1), Bujnicki JM(2).

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, PL-02-109 Warsaw,

Poland. (2)Laboratory of Bioinformatics and Protein Engineering, International

Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, PL-02-109

Warsaw, Poland Bioinformatics Laboratory, Institute of Molecular Biology and

Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89,

PL-61-614 Poznan, Poland iamb@genesilico.pl.

RNA function in many biological processes depends on the formation of

three-dimensional (3D) structures. However, RNA structure is difficult to

determine experimentally, which has prompted the development of predictive

computational methods. Here, we introduce a user-friendly online interface for

modeling RNA 3D structures using SimRNA, a method that uses a coarse-grained

representation of RNA molecules, utilizes the Monte Carlo method to sample the

conformational space, and relies on a statistical potential to describe the

interactions in the folding process. SimRNAweb makes SimRNA accessible to users

who do not normally use high performance computational facilities or are

unfamiliar with using the command line tools. The simplest input consists of an

RNA sequence to fold RNA de novo. Alternatively, a user can provide a 3D

structure in the PDB format, for instance a preliminary model built with some

other technique, to jump-start the modeling close to the expected final outcome.

The user can optionally provide secondary structure and distance restraints, and

can freeze a part of the starting 3D structure. SimRNAweb can be used to model

single RNA sequences and RNA-RNA complexes (up to 52 chains). The webserver is

available at http://genesilico.pl/SimRNAweb.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw279

PMCID: PMC4987879

PMID: 27095203

209. Nucleic Acids Res. 2016 Jul 8;44(W1):W242-5. doi: 10.1093/nar/gkw290. Epub 2016

Apr 19.

Interactive tree of life (iTOL) v3: an online tool for the display and annotation

of phylogenetic and other trees.

Letunic I(1), Bork P(2).

Author information:

(1)Biobyte solutions GmbH, Bothestr 142, 69126 Heidelberg, Germany

letunic@biobyte.de. (2)European Molecular Biology Laboratory, Meyerhofstrasse 1,

69117 Heidelberg, Germany Max Delbrück Centre for Molecular Medicine, 13125

Berlin, Germany Department of Bioinformatics, Biocenter, University of Würzburg,

97074 Würzburg, Germany.

Interactive Tree Of Life (http://itol.embl.de) is a web-based tool for the

display, manipulation and annotation of phylogenetic trees. It is freely

available and open to everyone. The current version was completely redesigned and

rewritten, utilizing current web technologies for speedy and streamlined

processing. Numerous new features were introduced and several new data types are

now supported. Trees with up to 100,000 leaves can now be efficiently displayed.

Full interactive control over precise positioning of various annotation features

and an unlimited number of datasets allow the easy creation of complex tree

visualizations. iTOL 3 is the first tool which supports direct visualization of

the recently proposed phylogenetic placements format. Finally, iTOL's account

system has been redesigned to simplify the management of trees in user-defined

workspaces and projects, as it is heavily used and currently handles already more

than 500,000 trees from more than 10,000 individual users.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw290

PMCID: PMC4987883

PMID: 27095192

210. Nucleic Acids Res. 2016 Jul 8;44(W1):W232-5. doi: 10.1093/nar/gkw256. Epub 2016

Apr 15.

W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis.

Trifinopoulos J(1), Nguyen LT(1), von Haeseler A(2), Minh BQ(3).

Author information:

(1)Center for Integrative Bioinformatics, Max F. Perutz Laboratories, University

of Vienna, Medical University of Vienna, 1030 Vienna, Austria. (2)Center for

Integrative Bioinformatics, Max F. Perutz Laboratories, University of Vienna,

Medical University of Vienna, 1030 Vienna, Austria Bioinformatics and

Computational Biology, Faculty of Computer Science, University of Vienna, 1090

Vienna, Austria. (3)Center for Integrative Bioinformatics, Max F. Perutz

Laboratories, University of Vienna, Medical University of Vienna, 1030 Vienna,

Austria minh.bui@univie.ac.at.

This article presents W-IQ-TREE, an intuitive and user-friendly web interface and

server for IQ-TREE, an efficient phylogenetic software for maximum likelihood

analysis. W-IQ-TREE supports multiple sequence types (DNA, protein, codon, binary

and morphology) in common alignment formats and a wide range of evolutionary

models including mixture and partition models. W-IQ-TREE performs fast model

selection, partition scheme finding, efficient tree reconstruction, ultrafast

bootstrapping, branch tests, and tree topology tests. All computations are

conducted on a dedicated computer cluster and the users receive the results via

URL or email. W-IQ-TREE is available at http://iqtree.cibiv.univie.ac.at It is

free and open to all users and there is no login requirement.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw256

PMCID: PMC4987875

PMID: 27084950

211. Nucleic Acids Res. 2016 Jul 8;44(W1):W201-4. doi: 10.1093/nar/gkw253. Epub 2016

Apr 15.

MBROLE 2.0-functional enrichment of chemical compounds.

López-Ibáñez J(1), Pazos F(1), Chagoyen M(2).

Author information:

(1)Computational Systems Biology Group, National Centre for Biotechnology

(CNB-CSIC), Madrid 28049, Spain. (2)Computational Systems Biology Group, National

Centre for Biotechnology (CNB-CSIC), Madrid 28049, Spain

monica.chagoyen@cnb.csic.es.

Metabolites Biological Role (MBROLE) is a server that performs functional

enrichment analysis of a list of chemical compounds derived from a metabolomics

experiment, which allows this list to be interpreted in biological terms. Since

its release in 2011, MBROLE has been used by different groups worldwide to

analyse metabolomics experiments from a variety of organisms. Here we present the

latest version of the system, MBROLE2, accessible at

http://csbg.cnb.csic.es/mbrole2 MBROLE2 has been supplemented with 10 databases

not available in the previous version, which allow analysis over a larger, richer

set of vocabularies including metabolite-protein and drug-protein interactions.

This new version performs automatic conversion of compound identifiers from

different databases, thus simplifying usage. In addition, the user interface has

been redesigned to generate an interactive, more intuitive representation of the

results.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw253

PMCID: PMC4987872

PMID: 27084944

212. Nucleic Acids Res. 2016 Jul 8;44(W1):W575-80. doi: 10.1093/nar/gkw254. Epub 2016

Apr 15.

MAGIC-web: a platform for untargeted and targeted N-linked glycoprotein

identification.

Lih TM(1), Choong WK(2), Chen CC(3), Cheng CW(1), Lin HN(2), Chen CT(2), Chang

HY(1), Hsu WL(2), Sung TY(4).

Author information:

(1)Bioinformatics Program, Taiwan International Graduate Program, Academia

Sinica, Taipei 11529, Taiwan Institute of Information Science, Academia Sinica,

Taipei 11529, Taiwan Institute of Biomedical Informatics, National Yang-Ming

University, Taipei 11221, Taiwan. (2)Institute of Information Science, Academia

Sinica, Taipei 11529, Taiwan. (3)Genomics Research Center, Academia Sinica,

Taipei 11529, Taiwan Department of Chemistry, National Taiwan University, Taipei

10617, Taiwan. (4)Institute of Information Science, Academia Sinica, Taipei

11529, Taiwan tsung@iis.sinica.edu.tw.

MAGIC-web is the first web server, to the best of our knowledge, that performs

both untargeted and targeted analyses of mass spectrometry-based glycoproteomics

data for site-specific N-linked glycoprotein identification. The first two

modules, MAGIC and MAGIC+, are designed for untargeted and targeted analysis,

respectively. MAGIC is implemented with our previously proposed novel Y1-ion

pattern matching method, which adequately detects Y1- and Y0-ion without prior

information of proteins and glycans, and then generates in silico MS(2) spectra

that serve as input to a database search engine (e.g. Mascot) to search against a

large-scale protein sequence database. On top of that, the newly implemented

MAGIC+ allows users to determine glycopeptide sequences using their own protein

sequence file. The third module, Reports Integrator, provides the service of

combining protein identification results from Mascot and glycan-related

information from MAGIC-web to generate a complete site-specific protein-glycan

summary report. The last module, Glycan Search, is designed for the users who are

interested in finding possible glycan structures with specific numbers and types

of monosaccharides. The results from MAGIC, MAGIC+ and Reports Integrator can be

downloaded via provided links whereas the annotated spectra and glycan structures

can be visualized in the browser. MAGIC-web is accessible from

http://ms.iis.sinica.edu.tw/MAGIC-web/index.html.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw254

PMCID: PMC4987873

PMID: 27084943

213. Nucleic Acids Res. 2016 Jul 8;44(W1):W562-7. doi: 10.1093/nar/gkw251. Epub 2016

Apr 15.

RBscore&NBench: a high-level web server for nucleic acid binding residues

prediction with a large-scale benchmarking database.

Miao Z(1), Westhof E(2).

Author information:

(1)Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de

biologie moléculaire et cellulaire du CNRS, 15 Rue Descartes 67000 Strasbourg,

France z.miao@ibmc-cnrs.unistra.fr. (2)Architecture et Réactivité de l'ARN,

Université de Strasbourg, Institut de biologie moléculaire et cellulaire du CNRS,

15 Rue Descartes 67000 Strasbourg, France e.westhof@ibmc-cnrs.unistra.fr.

RBscore&NBench combines a web server, RBscore and a database, NBench. RBscore

predicts RNA-/DNA-binding residues in proteins and visualizes the prediction

scores and features on protein structures. The scoring scheme of RBscore directly

links feature values to nucleic acid binding probabilities and illustrates the

nucleic acid binding energy funnel on the protein surface. To avoid dataset,

binding site definition and assessment metric biases, we compared RBscore with 18

web servers and 3 stand-alone programs on 41 datasets, which demonstrated the

high and stable accuracy of RBscore. A comprehensive comparison led us to develop

a benchmark database named NBench. The web server is available on:

http://ahsoka.u-strasbg.fr/rbscorenbench/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw251

PMCID: PMC4987871

PMID: 27084939

214. Nucleic Acids Res. 2016 Jul 8;44(W1):W581-6. doi: 10.1093/nar/gkw211. Epub 2016

Apr 15.

DeepBlue epigenomic data server: programmatic data retrieval and analysis of

epigenome region sets.

Albrecht F(1), List M(2), Bock C(3), Lengauer T(2).

Author information:

(1)Max Planck Institute for Informatics, 66123 Saarbrücken, Germany Graduate

School of Computer Science, Saarland University, 66123 Saarbrücken, Germany

felipe.albrecht@mpi-inf.mpg.de. (2)Max Planck Institute for Informatics, 66123

Saarbrücken, Germany. (3)Max Planck Institute for Informatics, 66123 Saarbrücken,

Germany CeMM Research Center for Molecular Medicine of the Austrian Academy of

Sciences, 1090 Vienna, Austria Department of Laboratory Medicine, Medical

University of Vienna, 1090 Vienna, Austria.

Large amounts of epigenomic data are generated under the umbrella of the

International Human Epigenome Consortium, which aims to establish 1000 reference

epigenomes within the next few years. These data have the potential to unravel

the complexity of epigenomic regulation. However, their effective use is hindered

by the lack of flexible and easy-to-use methods for data retrieval. Extracting

region sets of interest is a cumbersome task that involves several manual steps:

identifying the relevant experiments, downloading the corresponding data files

and filtering the region sets of interest. Here we present the DeepBlue

Epigenomic Data Server, which streamlines epigenomic data analysis as well as

software development. DeepBlue provides a comprehensive programmatic interface

for finding, selecting, filtering, summarizing and downloading region sets. It

contains data from four major epigenome projects, namely ENCODE, ROADMAP,

BLUEPRINT and DEEP. DeepBlue comes with a user manual, examples and a

well-documented application programming interface (API). The latter is accessed

via the XML-RPC protocol supported by many programming languages. To demonstrate

usage of the API and to enable convenient data retrieval for non-programmers, we

offer an optional web interface. DeepBlue can be openly accessed at

http://deepblue.mpi-inf.mpg.de.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw211

PMCID: PMC4987868

PMID: 27084938

215. Sci Rep. 2016 Jul 7;6:29575. doi: 10.1038/srep29575.

mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity

in genetic disease and emergence of drug resistance.

Pires DE(1,)(2), Blundell TL(1), Ascher DB(1).

Author information:

(1)Department of Biochemistry, Sanger Building, University of Cambridge, 80

Tennis Court Road, Cambridge, CB2 1GA, UK. (2)Centro de Pesquisas René Rachou,

Fundação Oswaldo Cruz, Avenida Augusto de Lima 1715, Belo Horizonte, 30190-002,

Brazil.

The ability to predict how a mutation affects ligand binding is an essential step

in understanding, anticipating and improving the design of new treatments for

drug resistance, and in understanding genetic diseases. Here we present mCSM-lig,

a structure-guided computational approach for quantifying the effects of

single-point missense mutations on affinities of small molecules for proteins.

mCSM-lig uses graph-based signatures to represent the wild-type environment of

mutations, and small-molecule chemical features and changes in protein stability

as evidence to train a predictive model using a representative set of

protein-ligand complexes from the Platinum database. We show our method provides

a very good correlation with experimental data (up to ρ = 0.67) and is effective

in predicting a range of chemotherapeutic, antiviral and antibiotic resistance

mutations, providing useful insights for genotypic screening and to guide drug

development. mCSM-lig also provides insights into understanding Mendelian disease

mutations and as a tool for guiding protein design. mCSM-lig is freely available

as a web server at http://structure.bioc.cam.ac.uk/mcsm\_lig.

DOI: 10.1038/srep29575

PMCID: PMC4935856

PMID: 27384129

216. J Comput Chem. 2016 Jul 5;37(18):1740-5. doi: 10.1002/jcc.24392. Epub 2016 May 8.

Ecoupling server: A tool to compute and analyze electronic couplings.

Cabeza de Vaca I(1,)(2), Acebes S(1), Guallar V(1,)(3).

Author information:

(1)Joint BSC-IRB Research Program in Computational Biology, Barcelona

Supercomputing Center, Life Science Department, Electronic and Atomic Modeling

Group, Nexus II, C/Jordi Girona, 29, Barcelona, 08034, Spain. (2)Department of

Chemistry, Yale University, New Haven, Connecticut, 06520-8107. (3)Institució

Catalana De Recerca I Estudis Avançats (ICREA), Barcelona, Spain.

Electron transfer processes are often studied through the evaluation and analysis

of the electronic coupling (EC). Since most standard QM codes do not provide

readily such a measure, additional, and user-friendly tools to compute and

analyze electronic coupling from external wave functions will be of high value.

The first server to provide a friendly interface for evaluation and analysis of

electronic couplings under two different approximations (FDC and GMH) is

presented in this communication. Ecoupling server accepts inputs from common QM

and QM/MM software and provides useful plots to understand and analyze the

results easily. The web server has been implemented in CGI-python using Apache

and it is accessible at http://ecouplingserver.bsc.es. Ecoupling server is free

and open to all users without login. © 2016 Wiley Periodicals, Inc.

© 2016 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24392

PMID: 27157013

217. J Comput Chem. 2016 Jul 5;37(18):1734-9. doi: 10.1002/jcc.24380. Epub 2016 Apr

13.

SPOT-Ligand: Fast and effective structure-based virtual screening by binding

homology search according to ligand and receptor similarity.

Yang Y(1), Zhan J(1), Zhou Y(1).

Author information:

(1)Institute for Glycomics and School of Information and Communication

Technology, Griffith University, Parklands Dr, Southport, QLD, 4222, Australia.

Structure-based virtual screening usually involves docking of a library of

chemical compounds onto the functional pocket of the target receptor so as to

discover novel classes of ligands. However, the overall success rate remains low

and screening a large library is computationally intensive. An alternative to

this "ab initio" approach is virtual screening by binding homology search. In

this approach, potential ligands are predicted based on similar interaction pairs

(similarity in receptors and ligands). SPOT-Ligand is an approach that integrates

ligand similarity by Tanimoto coefficient and receptor similarity by protein

structure alignment program SPalign. The method was found to yield a consistent

performance in DUD and DUD-E docking benchmarks even if model structures were

employed. It improves over docking methods (DOCK6 and AUTODOCK Vina) and has a

performance comparable to or better than other binding-homology methods (FINDsite

and PoLi) with higher computational efficiency. The server is available at

http://sparks-lab.org. © 2016 Wiley Periodicals, Inc.

© 2016 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24380

PMID: 27074979

218. Amino Acids. 2016 Jul;48(7):1655-65. doi: 10.1007/s00726-016-2226-z. Epub 2016

Apr 13.

SPAR: a random forest-based predictor for self-interacting proteins with

fine-grained domain information.

Liu X(1), Yang S(1), Li C(2), Zhang Z(3), Song J(4,)(5,)(6).

Author information:

(1)State Key Laboratory of Agrobiotechnology, College of Biological Sciences,

China Agricultural University, Beijing, 100193, China. (2)Infection and Immunity

Program, Biomedicine Discovery Institute and Department of Biochemistry and

Molecular Biology, Monash University, Melbourne, VIC, 3800, Australia. (3)State

Key Laboratory of Agrobiotechnology, College of Biological Sciences, China

Agricultural University, Beijing, 100193, China. zidingzhang@cau.edu.cn.

(4)Infection and Immunity Program, Biomedicine Discovery Institute and Department

of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC, 3800,

Australia. Jiangning.Song@monash.edu. (5)Monash Centre for Data Science, Faculty

of Information Technology, Monash University, Melbourne, VIC, 3800, Australia.

Jiangning.Song@monash.edu. (6)National Engineering Laboratory for Industrial

Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute

of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, 300308, China.

Jiangning.Song@monash.edu.

Protein self-interaction, i.e. the interaction between two or more identical

proteins expressed by one gene, plays an important role in the regulation of

cellular functions. Considering the limitations of experimental self-interaction

identification, it is necessary to design specific bioinformatics tools for

self-interacting protein (SIP) prediction from protein sequence information. In

this study, we proposed an improved computational approach for SIP prediction,

termed SPAR (Self-interacting Protein Analysis serveR). Firstly, we developed an

improved encoding scheme named critical residues substitution (CRS), in which the

fine-grained domain-domain interaction information was taken into account. Then,

by employing the Random Forest algorithm, the performance of CRS was evaluated

and compared with several other encoding schemes commonly used for sequence-based

protein-protein interaction prediction. Through the tenfold cross-validation

tests on a balanced training dataset, CRS performed the best, with the average

accuracy up to 72.01 %. We further integrated CRS with other encoding schemes and

identified the most important features using the mRMR (the minimum redundancy

maximum relevance) feature selection method. Our SPAR model with selected

features achieved an average accuracy of 92.09 % on the human-independent test

set (the ratio of positives to negatives was about 1:11). Besides, we also

evaluated the performance of SPAR on an independent yeast test set (the ratio of

positives to negatives was about 1:8) and obtained an average accuracy of

76.96 %. The results demonstrate that SPAR is capable of achieving a reasonable

performance in cross-species application. The SPAR server is freely available for

academic use at http://systbio.cau.edu.cn/zzdlab/spar/ .

DOI: 10.1007/s00726-016-2226-z

PMID: 27074717

219. Bioinformatics. 2016 Jul 1;32(13):2067-8. doi: 10.1093/bioinformatics/btw102.

Epub 2016 Feb 26.

NMRPro: an integrated web component for interactive processing and visualization

of NMR spectra.

Mohamed A(1), Nguyen CH(1), Mamitsuka H(1).

Author information:

(1)Bioinformatics Center, Institute for Chemical Research, Kyoto University,

Kyoto, Japan.

The popularity of using NMR spectroscopy in metabolomics and natural products has

driven the development of an array of NMR spectral analysis tools and databases.

Particularly, web applications are well used recently because they are

platform-independent and easy to extend through reusable web components.

Currently available web applications provide the analysis of NMR spectra.

However, they still lack the necessary processing and interactive visualization

functionalities. To overcome these limitations, we present NMRPro, a web

component that can be easily incorporated into current web applications, enabling

easy-to-use online interactive processing and visualization. NMRPro integrates

server-side processing with client-side interactive visualization through three

parts: a python package to efficiently process large NMR datasets on the

server-side, a Django App managing server-client interaction, and SpecdrawJS for

client-side interactive visualization.AVAILABILITY AND IMPLEMENTATION: Demo and

installation instructions are available at http://mamitsukalab.org/tools/nmrpro/

CONTACT: mohamed@kuicr.kyoto-u.ac.jp

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw102

PMID: 27153725

220. Bioinformatics. 2016 Jul 1;32(13):2017-23. doi: 10.1093/bioinformatics/btw103.

Epub 2016 Feb 24.

Cas-Database: web-based genome-wide guide RNA library design for gene knockout

screens using CRISPR-Cas9.

Park J(1), Kim JS(2), Bae S(3).

Author information:

(1)Department of Chemistry, Hanyang University, Seoul 133-791, South Korea.

(2)Center for Genome Engineering, Institute for Basic Science, Seoul 151-747,

South Korea Department of Chemistry, Seoul National University, Seoul 151-747,

South Korea. (3)Department of Chemistry, Hanyang University, Seoul 133-791, South

Korea Institute for Materials Design, Hanyang University, Seoul 133-791, South

Korea.

MOTIVATION: CRISPR-derived RNA guided endonucleases (RGENs) have been widely used

for both gene knockout and knock-in at the level of single or multiple genes.

RGENs are now available for forward genetic screens at genome scale, but single

guide RNA (sgRNA) selection at this scale is difficult.

RESULTS: We develop an online tool, Cas-Database, a genome-wide gRNA library

design tool for Cas9 nucleases from Streptococcus pyogenes (SpCas9). With an

easy-to-use web interface, Cas-Database allows users to select optimal target

sequences simply by changing the filtering conditions. Furthermore, it provides a

powerful way to select multiple optimal target sequences from thousands of genes

at once for the creation of a genome-wide library. Cas-Database also provides a

web application programming interface (web API) for advanced bioinformatics

users.

AVAILABILITY AND IMPLEMENTATION: Free access at

http://www.rgenome.net/cas-database/

CONTACT: sangsubae@hanyang.ac.kr or jskim01@snu.ac.kr

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw103

PMCID: PMC4920116

PMID: 27153724

221. Bioinformatics. 2016 Jul 1;32(13):2053-5. doi: 10.1093/bioinformatics/btw122.

Epub 2016 Mar 7.

ORFanFinder: automated identification of taxonomically restricted orphan genes.

Ekstrom A(1), Yin Y(2).

Author information:

(1)Department of Computer Science. (2)Department of Biological Sciences,

Montgomery Hall 325A, Northern Illinois University, DeKalb, IL, USA.

MOTIVATION: Orphan genes, also known as ORFans, are newly evolved genes in a

genome that enable the organism to adapt to specific living environment. The gene

content of every sequenced genome can be classified into different age groups,

based on how widely/narrowly a gene's homologs are distributed in the context of

species taxonomy. Those having homologs restricted to organisms of particular

taxonomic ranks are classified as taxonomically restricted ORFans.

RESULTS: Implementing this idea, we have developed an open source program named

ORFanFinder and a free web server to allow automated classification of a genome's

gene content and identification of ORFans at different taxonomic ranks.

ORFanFinder and its web server will contribute to the comparative genomics field

by facilitating the study of the origin of new genes and the emergence of

lineage-specific traits in both prokaryotes and eukaryotes.

AVAILABILITY AND IMPLEMENTATION: http://cys.bios.niu.edu/orfanfinder

CONTACT: yyin@niu.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw122

PMCID: PMC4920126 [Available on 2017-07-01]

PMID: 27153690

222. Bioinformatics. 2016 Jul 1;32(13):2050-2. doi: 10.1093/bioinformatics/btw119.

Epub 2016 Mar 7.

Visual Omics Explorer (VOE): a cross-platform portal for interactive data

visualization.

Kim B(1), Ali T(1), Hosmer S(2), Krampis K(3).

Author information:

(1)Department of Biological Sciences Center for Translational and Basic Research

and Belfer Research Building, Hunter College of the City University of New York,

New York, NY, USA. (2)The Graduate Center, City University of New York, New York,

NY, USA. (3)Department of Biological Sciences Center for Translational and Basic

Research and Belfer Research Building, Hunter College of the City University of

New York, New York, NY, USA The Graduate Center, City University of New York, New

York, NY, USA Department of Physiology and Biophysics, Institute for

Computational Biomedicine, Weil Cornell Medical College, Cornell University, New

York, NY, USA.

MOTIVATION: Given the abundance of genome sequencing and omics data, an

opprtunity and challenge in bioinformatics relates to data mining and

visualization. The majority of current bioinformatics visualizations are

implemented either as multi-tier web server applications that require significant

maintenance effort, or as client software that presumes technical expertise for

installation. Here we present the Visual Omics Explorer (VOE), a cross-platform

data visualization portal that is implemented using only HTML and Javascript

code. VOE is a standalone software that can be loaded offline on the web browser

from a local copy of the code, or over the internet without any dependency other

than distributing the code through a file sharing service. VOE can interactively

display genomics, transcriptomics, epigenomics and metagenomics data stored

either locally or retrieved from cloud storage services, and runs on both desktop

computers and mobile devices.

AVAILABILITY AND IMPLEMENTATION: VOE is accessible at http://bcil.github.io/VOE/

CONTACT: agbiotec@gmail.com

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw119

PMID: 27153572

223. IEEE/ACM Trans Comput Biol Bioinform. 2016 Jul-Aug;13(4):706-18. doi:

10.1109/TCBB.2015.2474407. Epub 2015 Aug 28.

Mem-mEN: Predicting Multi-Functional Types of Membrane Proteins by Interpretable

Elastic Nets.

Wan S, Mak MW, Kung SY.

Membrane proteins play important roles in various biological processes within

organisms. Predicting the functional types of membrane proteins is indispensable

to the characterization of membrane proteins. Recent studies have extended to

predicting single- and multi-type membrane proteins. However, existing predictors

perform poorly and more importantly, they are often lack of interpretability. To

address these problems, this paper proposes an efficient predictor, namely

Mem-mEN, which can produce sparse and interpretable solutions for predicting

membrane proteins with single- and multi-label functional types. Given a query

membrane protein, its associated gene ontology (GO) information is retrieved by

searching a compact GO-term database with its homologous accession number, which

is subsequently classified by a multi-label elastic net (EN) classifier.

Experimental results show that Mem-mEN significantly outperforms existing

state-of-the-art membrane-protein predictors. Moreover, by using Mem-mEN, 338 out

of more than 7,900 GO terms are found to play more essential roles in determining

the functional types. Based on these 338 essential GO terms, Mem-mEN can not only

predict the functional type of a membrane protein, but also explain why it

belongs to that type. For the reader's convenience, the Mem-mEN server is

available online at http://bioinfo.eie.polyu.edu.hk/MemmENServer/.

DOI: 10.1109/TCBB.2015.2474407

PMID: 26336143

224. J Nat Sci Biol Med. 2016 Jul-Dec;7(2):124-6. doi: 10.4103/0976-9668.184696.

PreFRP: Prediction and visualization of fluctuation residues in proteins.

Saravanan KM(1), Senthil R(2).

Author information:

(1)Centre of Excellence in Bioinformatics, School of Biotechnology, Madurai

Kamaraj University, Madurai, India. (2)Department of Bioinformatics, School of

Biosciences, The Marudupandiyar Institutions (Affiliated to Bharathidasan

University), Thanjavur, Tamil Nadu, India.

AIM: The PreFRP web server extracts sequence and basic information of a protein

structure and groups amino acid residues in a protein into three important types

such as high, moderate, and weak fluctuating residues.

MATERIALS AND METHODS: The server takes a protein data bank file or an amino acid

sequence as input and prints the probability of amino acid residues to fluctuate.

The server also provides a link to Jmol, a molecular visualization program to

visualize the high, moderate, and weak fluctuating residues in three different

colors.

RESULTS: Prediction and visualization of fluctuating amino acid residues in

proteins may help to understand the complex three-dimensional structure of

proteins and may further help in docking and mutation experiments.

AVAILABILITY: The web server is freely accessible through the web page of the

author's institution http://www.mpi.edu.in/prefrp/link.html.

DOI: 10.4103/0976-9668.184696

PMCID: PMC4934099

PMID: 27433060

225. PLoS One. 2016 Jun 30;11(6):e0158680. doi: 10.1371/journal.pone.0158680.

eCollection 2016.

How to Direct the Edges of the Connectomes: Dynamics of the Consensus Connectomes

and the Development of the Connections in the Human Brain.

Kerepesi C(1,)(2), Szalkai B(1), Varga B(1), Grolmusz V(1,)(3).

Author information:

(1)PIT Bioinformatics Group, Eötvös University, H-1117 Budapest, Hungary.

(2)Institute for Computer Science and Control, H-1111 Budapest, Hungary.

(3)Uratim Ltd., H-1118 Budapest, Hungary.

The human braingraph or the connectome is the object of an intensive research

today. The advantage of the graph-approach to brain science is that the rich

structures, algorithms and definitions of graph theory can be applied to the

anatomical networks of the connections of the human brain. In these graphs, the

vertices correspond to the small (1-1.5 cm2) areas of the gray matter, and two

vertices are connected by an edge, if a diffusion-MRI based workflow finds fibers

of axons, running between those small gray matter areas in the white matter of

the brain. One main question of the field today is discovering the directions of

the connections between the small gray matter areas. In a previous work we have

reported the construction of the Budapest Reference Connectome Server

http://connectome.pitgroup.org from the data recorded in the Human Connectome

Project of the NIH. The server generates the consensus braingraph of 96 subjects

in Version 2, and of 418 subjects in Version 3, according to selectable

parameters. After the Budapest Reference Connectome Server had been published, we

recognized a surprising and unforeseen property of the server. The server can

generate the braingraph of connections that are present in at least k graphs out

of the 418, for any value of k = 1, 2, …, 418. When the value of k is changed

from k = 418 through 1 by moving a slider at the webserver from right to left,

certainly more and more edges appear in the consensus graph. The astonishing

observation is that the appearance of the new edges is not random: it is similar

to a growing shrub. We refer to this phenomenon as the Consensus Connectome

Dynamics. We hypothesize that this movement of the slider in the webserver may

copy the development of the connections in the human brain in the following

sense: the connections that are present in all subjects are the oldest ones, and

those that are present only in a decreasing fraction of the subjects are

gradually the newer connections in the individual brain development. An animation

on the phenomenon is available at https://youtu.be/yxlyudPaVUE. Based on this

observation and the related hypothesis, we can assign directions to some of the

edges of the connectome as follows: Let Gk + 1 denote the consensus connectome

where each edge is present in at least k+1 graphs, and let Gk denote the

consensus connectome where each edge is present in at least k graphs. Suppose

that vertex v is not connected to any other vertices in Gk+1, and becomes

connected to a vertex u in Gk, where u was connected to other vertices already in

Gk+1. Then we direct this (v, u) edge from v to u.

DOI: 10.1371/journal.pone.0158680

PMCID: PMC4928947

PMID: 27362431

226. PLoS One. 2016 Jun 29;11(6):e0158568. doi: 10.1371/journal.pone.0158568.

eCollection 2016.

Characterizing Blood Metabolomics Profiles Associated with Self-Reported Food

Intakes in Female Twins.

Pallister T(1), Jennings A(2), Mohney RP(3), Yarand D(1), Mangino M(1), Cassidy

A(2), MacGregor A(2), Spector TD(1), Menni C(1).

Author information:

(1)Department of Twin Research and Genetic Epidemiology, King's College London,

London, United Kingdom. (2)Department of Nutrition, Norwich Medical School,

University of East Anglia, Norwich, United Kingdom. (3)Metabolon, Inc., Durham,

United States of America.

Using dietary biomarkers in nutritional epidemiological studies may better

capture exposure and improve the level at which diet-disease associations can be

established and explored. Here, we aimed to identify and evaluate reproducibility

of novel biomarkers of reported habitual food intake using targeted and

non-targeted metabolomic blood profiling in a large twin cohort. Reported intakes

of 71 food groups, determined by FFQ, were assessed against 601 fasting blood

metabolites in over 3500 adult female twins from the TwinsUK cohort. For each

metabolite, linear regression analysis was undertaken in the discovery group

(excluding MZ twin pairs discordant [≥1 SD apart] for food group intake) with

each food group as a predictor adjusting for age, batch effects, BMI, family

relatedness and multiple testing (1.17x10-6 = 0.05/[71 food groups x 601 detected

metabolites]). Significant results were then replicated (non-targeted: P<0.05;

targeted: same direction) in the MZ discordant twin group and results from both

analyses meta-analyzed. We identified and replicated 180 significant associations

with 39 food groups (P<1.17x10-6), overall consisting of 106 different

metabolites (74 known and 32 unknown), including 73 novel associations. In

particular we identified trans-4-hydroxyproline as a potential marker of red meat

intake (0.075[0.009]; P = 1.08x10-17), ergothioneine as a marker of mushroom

consumption (0.181[0.019]; P = 5.93x10-22), and three potential markers of fruit

consumption (top association: apple and pears): including metabolites derived

from gut bacterial transformation of phenolic compounds, 3-phenylpropionate

(0.024[0.004]; P = 1.24x10-8) and indolepropionate (0.026[0.004]; P = 2.39x10-9),

and threitol (0.033[0.003]; P = 1.69x10-21). With the largest nutritional

metabolomics dataset to date, we have identified 73 novel candidate biomarkers of

food intake for potential use in nutritional epidemiological studies. We compiled

our findings into the DietMetab database

(http://www.twinsuk.ac.uk/dietmetab-data/), an online tool to investigate our top

associations.

DOI: 10.1371/journal.pone.0158568

PMCID: PMC4927065

PMID: 27355821

227. PLoS Comput Biol. 2016 Jun 23;12(6):e1004809. doi: 10.1371/journal.pcbi.1004809.

eCollection 2016.

QuIN: A Web Server for Querying and Visualizing Chromatin Interaction Networks.

Thibodeau A(1), Márquez EJ(2), Luo O(2), Ruan Y(2), Menghi F(2), Shin DG(1),

Stitzel ML(2,)(3), Vera-Licona P(3,)(4), Ucar D(2,)(3).

Author information:

(1)Department of Computer Science, University of Connecticut, Storrs,

Connecticut, United States of America. (2)The Jackson Laboratory for Genomic

Medicine, Farmington, Connecticut, United States of America. (3)Institute of

Systems Genomics, University of Connecticut Health Center, Farmington,

Connecticut, United States of America. (4)Center for Quantitative Medicine,

Department of Cell Biology, University of Connecticut Health Center, Farmington,

Connecticut, United States of America.

Recent studies of the human genome have indicated that regulatory elements (e.g.

promoters and enhancers) at distal genomic locations can interact with each other

via chromatin folding and affect gene expression levels. Genomic technologies for

mapping interactions between DNA regions, e.g., ChIA-PET and HiC, can generate

genome-wide maps of interactions between regulatory elements. These interaction

datasets are important resources to infer distal gene targets of non-coding

regulatory elements and to facilitate prioritization of critical loci for

important cellular functions. With the increasing diversity and complexity of

genomic information and public ontologies, making sense of these datasets demands

integrative and easy-to-use software tools. Moreover, network representation of

chromatin interaction maps enables effective data visualization, integration, and

mining. Currently, there is no software that can take full advantage of network

theory approaches for the analysis of chromatin interaction datasets. To fill

this gap, we developed a web-based application, QuIN, which enables: 1) building

and visualizing chromatin interaction networks, 2) annotating networks with

user-provided private and publicly available functional genomics and interaction

datasets, 3) querying network components based on gene name or chromosome

location, and 4) utilizing network based measures to identify and prioritize

critical regulatory targets and their direct and indirect

interactions.AVAILABILITY: QuIN's web server is available at http://quin.jax.org

QuIN is developed in Java and JavaScript, utilizing an Apache Tomcat web server

and MySQL database and the source code is available under the GPLV3 license

available on GitHub: https://github.com/UcarLab/QuIN/.

DOI: 10.1371/journal.pcbi.1004809

PMCID: PMC4919057

PMID: 27336171

228. PLoS One. 2016 Jun 23;11(6):e0158287. doi: 10.1371/journal.pone.0158287.

eCollection 2016.

Platelets Proteomic Profiles of Acute Ischemic Stroke Patients.

Cevik O(1,)(2), Baykal AT(3), Sener A(2).

Author information:

(1)Cumhuriyet University, Faculty of Pharmacy, Department of Biochemistry, Sivas,

Turkey. (2)Marmara University, Faculty of Pharmacy, Department of Biochemistry,

Istanbul, Turkey. (3)Acibadem University, School of Medicine, Department of

Medical Biochemistry, Istanbul, Turkey.

Platelets play a crucial role in the pathogenesis of stroke and antiplatelet

agents exist for its treatment and prevention. Through the use of LC-MS based

protein expression profiling, platelets from stroke patients were analyzed and

then correlated with the proteomic analyses results in the context of this

disease. This study was based on patients who post ischemic stroke were admitted

to hospital and had venous blood drawn within 24 hrs of the incidence. Label-free

protein expression analyses of the platelets' tryptic digest was performed in

triplicate on a UPLC-ESI-qTOF-MS/MS system and ProteinLynx Global Server (v2.5,

Waters) was used for tandem mass data extraction. The peptide sequences were

searched against the reviewed homo sapiens database (www.uniprot.org) and the

quantitation of protein variation was achieved through Progenesis LC-MS software

(V4.0, Nonlinear Dynamics). These Label-free differential proteomics analysis of

platelets ensured that 500 proteins were identified and 83 of these proteins were

found to be statistically significant. The differentially expressed proteins are

involved in various processes such as inflammatory response, cellular movement,

immune cell trafficking, cell-to-cell signaling and interaction, hematological

system development and function and nucleic acid metabolism. The expressions of

myeloperoxidase, arachidonate 12-Lipoxygenase and histidine-rich glycoprotein are

involved in cellular metabolic processes, crk-like protein and ras homolog gene

family member A involved in cell signaling with vitronectin, thrombospondin 1,

Integrin alpha 2b, and integrin beta 3 involved in cell adhesion. Apolipoprotein

H, immunoglobulin heavy constant gamma 1 and immunoglobulin heavy constant gamma

3 are involved in structural, apolipoprotein A-I, and

alpha-1-microglobulin/bikunin precursor is involved in transport, complement

component 3 and clusterin is involved in immunity proteins as has been discussed.

Our data provides an insight into the proteins that are involved in the

platelets' activation response during ischemic stroke. It could be argued that

this study lays the foundation for future mechanistic studies.

DOI: 10.1371/journal.pone.0158287

PMCID: PMC4919045

PMID: 27336623

229. Front Plant Sci. 2016 Jun 21;7:889. doi: 10.3389/fpls.2016.00889. eCollection

2016.

PlantAPA: A Portal for Visualization and Analysis of Alternative Polyadenylation

in Plants.

Wu X(1), Zhang Y(1), Li QQ(2).

Author information:

(1)Department of Automation, Xiamen University Xiamen, China. (2)Key Laboratory

of the Ministry of Education for Coastal and Wetland Ecosystems, College of the

Environment and Ecology, Xiamen UniversityXiamen, China; Graduate College of

Biomedical Sciences, Western University of Health SciencesPomona, CA, USA.

Alternative polyadenylation (APA) is an important layer of gene regulation that

produces mRNAs that have different 3' ends and/or encode diverse protein

isoforms. Up to 70% of annotated genes in plants undergo APA. Increasing numbers

of poly(A) sites collected in various plant species demand new methods and tools

to access and mine these data. We have created an open-access web service called

PlantAPA (http://bmi.xmu.edu.cn/plantapa) to visualize and analyze genome-wide

poly(A) sites in plants. PlantAPA provides various interactive and dynamic

graphics and seamlessly integrates a genome browser that can profile

heterogeneous cleavage sites and quantify expression patterns of poly(A) sites

across different conditions. Particularly, through PlantAPA, users can analyze

poly(A) sites in extended 3' UTR regions, intergenic regions, and ambiguous

regions owing to alternative transcription or RNA processing. In addition, it

also provides tools for analyzing poly(A) site selections, 3' UTR lengthening or

shortening, non-canonical APA site switching, and differential gene expression

between conditions, making it more powerful for the study of APA-mediated gene

expression regulation. More importantly, PlantAPA offers a bioinformatics

pipeline that allows users to upload their own short reads or ESTs for poly(A)

site extraction, enabling users to further explore poly(A) site selection using

stored PlantAPA poly(A) sites together with their own poly(A) site datasets. To

date, PlantAPA hosts the largest database of APA sites in plants, including Oryza

sativa, Arabidopsis thaliana, Medicago truncatula, and Chlamydomonas reinhardtii.

As a user-friendly web service, PlantAPA will be a valuable addition to the

community of biologists studying APA mechanisms and gene expression regulation in

plants.

DOI: 10.3389/fpls.2016.00889

PMCID: PMC4914594

PMID: 27446120

230. Sci Rep. 2016 Jun 20;6:28268. doi: 10.1038/srep28268.

ORION: a web server for protein fold recognition and structure prediction using

evolutionary hybrid profiles.

Ghouzam Y(1,)(2,)(3,)(4), Postic G(1,)(2,)(3,)(4), Guerin PE(1,)(2,)(3,)(4), de

Brevern AG(1,)(2,)(3,)(4), Gelly JC(1,)(2,)(3,)(4).

Author information:

(1)INSERM, U 1134, DSIMB, F-75739 Paris, France. (2)Univ. Paris Diderot, Sorbonne

Paris Cité, UMR\_S 1134, F-75739 Paris, France. (3)Institut National de la

Transfusion Sanguine (INTS), F-75739 Paris, France. (4)Laboratoire d'Excellence

GR-Ex, F-75739 Paris, France.

Protein structure prediction based on comparative modeling is the most efficient

way to produce structural models when it can be performed. ORION is a dedicated

webserver based on a new strategy that performs this task. The identification by

ORION of suitable templates is performed using an original profile-profile

approach that combines sequence and structure evolution information. Structure

evolution information is encoded into profiles using structural features, such as

solvent accessibility and local conformation -with Protein Blocks-, which give an

accurate description of the local protein structure. ORION has recently been

improved, increasing by 5% the quality of its results. The ORION web server

accepts a single protein sequence as input and searches homologous protein

structures within minutes. Various databases such as PDB, SCOP and HOMSTRAD can

be mined to find an appropriate structural template. For the modeling step, a

protein 3D structure can be directly obtained from the selected template by

MODELLER and displayed with global and local quality model estimation measures.

The sequence and the predicted structure of 4 examples from the CAMEO server and

a recent CASP11 target from the 'Hard' category (T0818-D1) are shown as pertinent

examples. Our web server is accessible at http://www.dsimb.inserm.fr/ORION/.

DOI: 10.1038/srep28268

PMCID: PMC4913311

PMID: 27319297

231. BMC Bioinformatics. 2016 Jun 17;17(1):242. doi: 10.1186/s12859-016-1124-4.

Detection and sequence/structure mapping of biophysical constraints to protein

variation in saturated mutational libraries and protein sequence alignments with

a dedicated server.

Abriata LA(1), Bovigny C(2,)(3), Dal Peraro M(2).

Author information:

(1)Laboratory for Biomolecular Modeling, Institute of Bioengineering, School of

Life Sciences, École Polytechnique Fédérale de Lausanne, and Swiss Institute of

Bioinformatics, AAB014 Station 19, Lausanne, 1015, Switzerland.

luciano.abriata@epfl.ch. (2)Laboratory for Biomolecular Modeling, Institute of

Bioengineering, School of Life Sciences, École Polytechnique Fédérale de

Lausanne, and Swiss Institute of Bioinformatics, AAB014 Station 19, Lausanne,

1015, Switzerland. (3)Present address: Molecular Modeling Group, Swiss Institute

of Bioinformatics, UNIL, Bâtiment Génopode, Lausanne, 1015, Switzerland.

Erratum in

BMC Bioinformatics. 2016 Oct 31;17 (1):439.

BACKGROUND: Protein variability can now be studied by measuring high-resolution

tolerance-to-substitution maps and fitness landscapes in saturated mutational

libraries. But these rich and expensive datasets are typically interpreted

coarsely, restricting detailed analyses to positions of extremely high or low

variability or dubbed important beforehand based on existing knowledge about

active sites, interaction surfaces, (de)stabilizing mutations, etc.

RESULTS: Our new webserver PsychoProt (freely available without registration at

http://psychoprot.epfl.ch or at

http://lucianoabriata.altervista.org/psychoprot/index.html ) helps to detect,

quantify, and sequence/structure map the biophysical and biochemical traits that

shape amino acid preferences throughout a protein as determined by

deep-sequencing of saturated mutational libraries or from large alignments of

naturally occurring variants.

DISCUSSION: We exemplify how PsychoProt helps to (i) unveil protein

structure-function relationships from experiments and from alignments that are

consistent with structures according to coevolution analysis, (ii) recall global

information about structural and functional features and identify hitherto

unknown constraints to variation in alignments, and (iii) point at different

sources of variation among related experimental datasets or between experimental

and alignment-based data. Remarkably, metabolic costs of the amino acids pose

strong constraints to variability at protein surfaces in nature but not in the

laboratory. This and other differences call for caution when extrapolating

results from in vitro experiments to natural scenarios in, for example, studies

of protein evolution.

CONCLUSION: We show through examples how PsychoProt can be a useful tool for the

broad communities of structural biology and molecular evolution, particularly for

studies about protein modeling, evolution and design.

DOI: 10.1186/s12859-016-1124-4

PMCID: PMC4912743

PMID: 27315797

232. Version 2. F1000Res. 2016 Jun 16 [revised 2016 Jul 20];5:1396. doi:

10.12688/f1000research.8798.2. eCollection 2016.

search.bioPreprint: a discovery tool for cutting edge, preprint biomedical

research articles.

Iwema CL(1), LaDue J(1), Zack A(1), Chattopadhyay A(1).

Author information:

(1)Health Sciences Library System, University of Pittsburgh, Pittsburgh, USA.

The time it takes for a completed manuscript to be published traditionally can be

extremely lengthy. Article publication delay, which occurs in part due to

constraints associated with peer review, can prevent the timely dissemination of

critical and actionable data associated with new information on rare diseases or

developing health concerns such as Zika virus. Preprint servers are open access

online repositories housing preprint research articles that enable authors (1) to

make their research immediately and freely available and (2) to receive

commentary and peer review prior to journal submission. There is a growing

movement of preprint advocates aiming to change the current journal publication

and peer review system, proposing that preprints catalyze biomedical discovery,

support career advancement, and improve scientific communication. While the

number of articles submitted to and hosted by preprint servers are gradually

increasing, there has been no simple way to identify biomedical research

published in a preprint format, as they are not typically indexed and are only

discoverable by directly searching the specific preprint server websites. To

address this issue, we created a search engine that quickly compiles preprints

from disparate host repositories and provides a one-stop search solution.

Additionally, we developed a web application that bolsters the discovery of

preprints by enabling each and every word or phrase appearing on any web site to

be integrated with articles from preprint servers. This tool, search.bioPreprint,

is publicly available at http://www.hsls.pitt.edu/resources/preprint.

DOI: 10.12688/f1000research.8798.2

PMCID: PMC4957174

PMID: 27508060

233. Front Microbiol. 2016 Jun 16;7:949. doi: 10.3389/fmicb.2016.00949. eCollection

2016.

Prediction of Biofilm Inhibiting Peptides: An In silico Approach.

Gupta S(1), Sharma AK(1), Jaiswal SK(1), Sharma VK(1).

Author information:

(1)Metagenomics and Systems Biology Group, Department of Biological Sciences,

Indian Institute of Science Education and Research Bhopal Bhopal, India.

Approximately 75% of microbial infections found in humans are caused by microbial

biofilms. These biofilms are resistant to host immune system and most of the

currently available antibiotics. Small peptides are extensively studied for their

role as anti-microbial peptides, however, only a limited studies have shown their

potential as inhibitors of biofilm. Therefore, to develop a unique computational

method aimed at the prediction of biofilm inhibiting peptides, the experimentally

validated biofilm inhibiting peptides sequences were used to extract sequence

based features and to identify unique sequence motifs. Biofilm inhibiting

peptides were observed to be abundant in positively charged and aromatic amino

acids, and also showed selective abundance of some dipeptides and sequence

motifs. These individual sequence based features were utilized to construct

Support Vector Machine-based prediction models and additionally by including

sequence motifs information, the hybrid models were constructed. Using 10-fold

cross validation, the hybrid model displayed the accuracy and Matthews

Correlation Coefficient (MCC) of 97.83% and 0.87, respectively. On the validation

dataset, the hybrid model showed the accuracy and MCC value of 97.19% and 0.84,

respectively. The validated model and other tools developed for the prediction of

biofilm inhibiting peptides are available freely as web server at

http://metagenomics.iiserb.ac.in/biofin/ and

http://metabiosys.iiserb.ac.in/biofin/.

DOI: 10.3389/fmicb.2016.00949

PMCID: PMC4909740

PMID: 27379078

234. Sci Rep. 2016 Jun 16;6:28249. doi: 10.1038/srep28249.

GPS-Lipid: a robust tool for the prediction of multiple lipid modification sites.

Xie Y(1), Zheng Y(1), Li H(1), Luo X(1), He Z(1), Cao S(1), Shi Y(1), Zhao

Q(1,)(2,)(3), Xue Y(4), Zuo Z(1,)(2), Ren J(1,)(2,)(3).

Author information:

(1)State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen

University, Guangzhou, Guangdong 510275, China. (2)Sun Yat-sen University Cancer

Center, State Key Laboratory of Oncology in South China, Collaborative Innovation

Center for Cancer Medicine, Guangzhou 510060, China. (3)Collaborative Innovation

Center of High Performance Computing, National University of Defense Technology,

Changsha 410073, China. (4)Department of Biomedical Engineering, College of Life

Science and Technology, Huazhong University of Science and Technology, Wuhan

430074, China.

As one of the most common post-translational modifications in eukaryotic cells,

lipid modification is an important mechanism for the regulation of variety

aspects of protein function. Over the last decades, three classes of lipid

modifications have been increasingly studied. The co-regulation of these

different lipid modifications is beginning to be noticed. However, due to the

lack of integrated bioinformatics resources, the studies of co-regulatory

mechanisms are still very limited. In this work, we developed a tool called

GPS-Lipid for the prediction of four classes of lipid modifications by

integrating the Particle Swarm Optimization with an aging leader and challengers

(ALC-PSO) algorithm. GPS-Lipid was proven to be evidently superior to other

similar tools. To facilitate the research of lipid modification, we hosted a

publicly available web server at http://lipid.biocuckoo.org with not only the

implementation of GPS-Lipid, but also an integrative database and visualization

tool. We performed a systematic analysis of the co-regulatory mechanism between

different lipid modifications with GPS-Lipid. The results demonstrated that the

proximal dual-lipid modifications among palmitoylation, myristoylation and

prenylation are key mechanism for regulating various protein functions. In

conclusion, GPS-lipid is expected to serve as useful resource for the research on

lipid modifications, especially on their co-regulation.

DOI: 10.1038/srep28249

PMCID: PMC4910163

PMID: 27306108

235. Bioinformatics. 2016 Jun 15;32(12):i360-i368. doi: 10.1093/bioinformatics/btw265.

RNAiFold2T: Constraint Programming design of thermo-IRES switches.

Garcia-Martin JA(1), Dotu I(2), Fernandez-Chamorro J(3), Lozano G(3), Ramajo

J(3), Martinez-Salas E(3), Clote P(1).

Author information:

(1)Biology Department, Boston College, Chestnut Hill, MA 02467, USA.

(2)Department of Experimental and Health Sciences, Research Programme on

Biomedical Informatics (GRIB), Universitat Pompeu Fabra. Dr. Aiguader 88,

Barcelona, Spain. (3)Centro de Biologia Molecular Severo Ochoa, Consejo Superior

de Investigaciones Cientificas-Universidad Autonoma de Madrid, 28049 Madrid,

Spain.

MOTIVATION: RNA thermometers (RNATs) are cis-regulatory elements that change

secondary structure upon temperature shift. Often involved in the regulation of

heat shock, cold shock and virulence genes, RNATs constitute an interesting

potential resource in synthetic biology, where engineered RNATs could prove to be

useful tools in biosensors and conditional gene regulation.

RESULTS: Solving the 2-temperature inverse folding problem is critical for RNAT

engineering. Here we introduce RNAiFold2T, the first Constraint Programming (CP)

and Large Neighborhood Search (LNS) algorithms to solve this problem.

Benchmarking tests of RNAiFold2T against existent programs (adaptive walk and

genetic algorithm) inverse folding show that our software generates two orders of

magnitude more solutions, thus allowing ample exploration of the space of

solutions. Subsequently, solutions can be prioritized by computing various

measures, including probability of target structure in the ensemble, melting

temperature, etc. Using this strategy, we rationally designed two thermosensor

internal ribosome entry site (thermo-IRES) elements, whose normalized

cap-independent translation efficiency is approximately 50% greater at 42 °C than

30 °C, when tested in reticulocyte lysates. Translation efficiency is lower than

that of the wild-type IRES element, which on the other hand is fully resistant to

temperature shift-up. This appears to be the first purely computational design of

functional RNA thermoswitches, and certainly the first purely computational

design of functional thermo-IRES elements.

AVAILABILITY: RNAiFold2T is publicly available as part of the new release

RNAiFold3.0 at https://github.com/clotelab/RNAiFold and

http://bioinformatics.bc.edu/clotelab/RNAiFold, which latter has a web server as

well. The software is written in C ++ and uses OR-Tools CP search engine.

CONTACT: clote@bc.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw265

PMCID: PMC4908349

PMID: 27307638

236. Bioinformatics. 2016 Jun 15;32(12):1885-7. doi: 10.1093/bioinformatics/btw082.

Epub 2016 Feb 15.

Multiple structure single parameter: analysis of a single protein nano

environment descriptor characterizing a shared loci on structurally aligned

proteins.

Salim JA(1), Borro L(1), Mazoni I(2), Yano I(2), Jardine JG(2), Neshich G(2).

Author information:

(1)EMBRAPA Agriculture Informatics, Computational Biology Research Group,

Brazilian Agricultural Research Corporation, Campinas 13083-886, Brazil and State

University of Campinas/UNICAMP, Campinas, 13081-970, Brazil. (2)EMBRAPA

Agriculture Informatics, Computational Biology Research Group, Brazilian

Agricultural Research Corporation, Campinas 13083-886, Brazil and.

MOTIVATION: A graphical representation of physicochemical and structural

descriptors attributed to amino acid residues occupying the same topological

position in different, structurally aligned proteins can provide a more intuitive

way to associate possible functional implications to identified variations in

structural characteristics. This could be achieved by observing selected

characteristics of amino acids and of their corresponding nano environments,

described by the numerical value of matching descriptor. For this purpose, a

web-based tool called multiple structure single parameter (MSSP) was developed

and here presented.

RESULTS: MSSP produces a two-dimensional plot of a single protein descriptor for

a number of structurally aligned protein chains. From a total of 150 protein

descriptors available in MSSP, selected of >1500 parameters stored in the STING

database, it is possible to create easily readable and highly informative

XY-plots, where X-axis contains the amino acid position in the multiple

structural alignment, and Y-axis contains the descriptor's numerical values for

each aligned structure. To illustrate one of possible MSSP contributions to the

investigation of changes in physicochemical and structural properties of mutants,

comparing them with the cognate wild-type structure, the oncogenic mutation of

M918T in RET kinase is presented. The comparative analysis of wild-type and

mutant structures shows great changes in their electrostatic potential. These

variations are easily depicted at the MSSP-generated XY-plot.

AVAILABILITY AND IMPLEMENTATION: The web server is freely available at

http://www.cbi.cnptia.embrapa.br/SMS/STINGm/MPA/index.html Web server implemented

in Perl, Java and JavaScript and JMol or Protein Viewer as structure visualizers.

CONTACT: goran.neshich@embrapa.br or gneshich@gmail.com

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw082

PMID: 27153716

237. Bioinformation. 2016 Jun 15;12(3):231-232. doi: 10.6026/97320630012231.

eCollection 2016.

Genes2GO: A web application for querying gene sets for specific GO terms.

Chawla K(1), Kuiper M(1).

Author information:

(1)Department of Biology, Norwegian University of Science and Technology (NTNU),

Hogskoleringen 5, 7491 Trondheim, Norway.

Gene ontology annotations have become an essential resource for biological

interpretations of experimental findings. The process of gathering basic

annotation information in tables that link gene sets with specific gene ontology

terms can be cumbersome, in particular if it requires above average computer

skills or bioinformatics expertise. We have therefore developed Genes2GO, an

intuitive R-based web application. Genes2GO uses the biomaRt package of

Bioconductor in order to retrieve custom sets of gene ontology annotations for

any list of genes from organisms covered by the Ensembl database. Genes2GO

produces a binary matrix file, indicating for each gene the presence or absence

of specific annotations for a gene. It should be noted that other GO tools do not

offer this user-friendly access to annotations.AVAILABILITY: Genes2GO is freely

available and listed under

http://www.semantic-systems-biology.org/tools/externaltools/.

DOI: 10.6026/97320630012231

PMCID: PMC5267968

PMID: 28149059

238. Bioinform Biol Insights. 2016 Jun 14;10:73-80. doi: 10.4137/BBI.S38423.

eCollection 2016.

PDBparam: Online Resource for Computing Structural Parameters of Proteins.

Nagarajan R(1), Archana A(1), Thangakani AM(2), Jemimah S(1), Velmurugan D(2),

Gromiha MM(1).

Author information:

(1)Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences,

Indian Institute of Technology Madras, Chennai, India. (2)CAS in Crystallography

and Biophysics, University of Madras, Chennai, India.; Bioinformatics

Infrastructure Facility, University of Madras, Chennai, India.

Understanding the structure-function relationship in proteins is a longstanding

goal in molecular and computational biology. The development of structure-based

parameters has helped to relate the structure with the function of a protein.

Although several structural features have been reported in the literature, no

single server can calculate a wide-ranging set of structure-based features from

protein three-dimensional structures. In this work, we have developed a web-based

tool, PDBparam, for computing more than 50 structure-based features for any given

protein structure. These features are classified into four major categories: (i)

interresidue interactions, which include short-, medium-, and long-range

interactions, contact order, long-range order, total contact distance, contact

number, and multiple contact index, (ii) secondary structure propensities such as

α-helical propensity, β-sheet propensity, and propensity of amino acids to exist

at various positions of α-helix and amino acid compositions in high B-value

regions, (iii) physicochemical properties containing ionic interactions, hydrogen

bond interactions, hydrophobic interactions, disulfide interactions, aromatic

interactions, surrounding hydrophobicity, and buriedness, and (iv) identification

of binding site residues in protein-protein, protein-nucleic acid, and

protein-ligand complexes. The server can be freely accessed at

http://www.iitm.ac.in/bioinfo/pdbparam/. We suggest the use of PDBparam as an

effective tool for analyzing protein structures.

DOI: 10.4137/BBI.S38423

PMCID: PMC4909059

PMID: 27330281

239. J Transl Med. 2016 Jun 14;14(1):178. doi: 10.1186/s12967-016-0928-3.

ProInflam: a webserver for the prediction of proinflammatory antigenicity of

peptides and proteins.

Gupta S(1), Madhu MK(1), Sharma AK(1), Sharma VK(2).

Author information:

(1)Metagenomics and Systems Biology Group, Department of Biological Sciences,

Indian Institute of Science Education and Research Bhopal, Bhopal, Madhya

Pradesh, India. (2)Metagenomics and Systems Biology Group, Department of

Biological Sciences, Indian Institute of Science Education and Research Bhopal,

Bhopal, Madhya Pradesh, India. vineetks@iiserb.ac.in.

BACKGROUND: Proinflammatory immune response involves a complex series of

molecular events leading to inflammatory reaction at a site, which enables host

to combat plurality of infectious agents. It can be initiated by specific stimuli

such as viral, bacterial, parasitic or allergenic antigens, or by non-specific

stimuli such as LPS. On counter with such antigens, the complex interaction of

antigen presenting cells, T cells and inflammatory mediators like IL1α, IL1β,

TNFα, IL12, IL18 and IL23 lead to proinflammatory immune response and further

clearance of infection. In this study, we have tried to establish a relation

between amino acid sequence of antigen and induction of proinflammatory response.

RESULTS: A total of 729 experimentally-validated proinflammatory and 171

non-proinflammatory epitopes were obtained from IEDB database. The A, F, I, L and

V amino acids and AF, FA, FF, PF, IV, IN dipeptides were observed as preferred

residues in proinflammatory epitopes. Using the compositional and motif-based

features of proinflammatory and non-proinflammatory epitopes, we have developed

machine learning-based models for prediction of proinflammatory response of

peptides. The hybrid of motifs and dipeptide-based features displayed best

performance with MCC = 0.58 and an accuracy of 87.6 %.

CONCLUSION: The amino acid sequence-based features of peptides were used to

develop a machine learning-based prediction tool for the prediction of

proinflammatory epitopes. This is a unique tool for the computational

identification of proinflammatory peptide antigen/candidates and provides leads

for experimental validations. The prediction model and tools for epitope mapping

and similarity search are provided as a comprehensive web server which is freely

available at http://metagenomics.iiserb.ac.in/proinflam/ and

http://metabiosys.iiserb.ac.in/proinflam/ .

DOI: 10.1186/s12967-016-0928-3

PMCID: PMC4908730

PMID: 27301453

240. Sci Rep. 2016 Jun 14;6:27930. doi: 10.1038/srep27930.

Antimicrobial Resistance Prediction in PATRIC and RAST.

Davis JJ(1,)(2), Boisvert S(3), Brettin T(1,)(2), Kenyon RW(4), Mao C(4), Olson

R(1,)(2), Overbeek R(2,)(5), Santerre J(6), Shukla M(1,)(2), Wattam AR(4), Will

R(4), Xia F(1,)(2), Stevens R(1,)(2,)(6).

Author information:

(1)University of Chicago, Computation Institute, 5735 South Ellis Avenue,

Chicago, IL 60637, USA. (2)Argonne National Laboratory, 9700 Cass Ave, Lemont, IL

60439, USA. (3)Gydle Inc. 101-1332 Chanoine Morel Quebec, QC, G1S, 4B4, Canada.

(4)Biocomplexity Institute of Virginia Tech, 1015 Life Science Cir, Blacksburg,

VA 24061, USA. (5)The Fellowship for Interpretation of Genomes, 15w155 81st St,

Burr Ridge, IL 60527, USA. (6)University of Chicago, Department of Computer

Science, Ryerson Physical Laboratory, 1100 E 58th St, Chicago, IL 60637, USA.

The emergence and spread of antimicrobial resistance (AMR) mechanisms in

bacterial pathogens, coupled with the dwindling number of effective antibiotics,

has created a global health crisis. Being able to identify the genetic mechanisms

of AMR and predict the resistance phenotypes of bacterial pathogens prior to

culturing could inform clinical decision-making and improve reaction time. At

PATRIC (http://patricbrc.org/), we have been collecting bacterial genomes with

AMR metadata for several years. In order to advance phenotype prediction and the

identification of genomic regions relating to AMR, we have updated the PATRIC FTP

server to enable access to genomes that are binned by their AMR phenotypes, as

well as metadata including minimum inhibitory concentrations. Using this

infrastructure, we custom built AdaBoost (adaptive boosting) machine learning

classifiers for identifying carbapenem resistance in Acinetobacter baumannii,

methicillin resistance in Staphylococcus aureus, and beta-lactam and

co-trimoxazole resistance in Streptococcus pneumoniae with accuracies ranging

from 88-99%. We also did this for isoniazid, kanamycin, ofloxacin, rifampicin,

and streptomycin resistance in Mycobacterium tuberculosis, achieving accuracies

ranging from 71-88%. This set of classifiers has been used to provide an initial

framework for species-specific AMR phenotype and genomic feature prediction in

the RAST and PATRIC annotation services.

DOI: 10.1038/srep27930

PMCID: PMC4906388

PMID: 27297683

241. Sci Rep. 2016 Jun 10;6:27653. doi: 10.1038/srep27653.

PDNAsite: Identification of DNA-binding Site from Protein Sequence by

Incorporating Spatial and Sequence Context.

Zhou J(1,)(2), Xu R(1,)(3), He Y(4), Lu Q(2), Wang H(1), Kong B(1).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong, China. (2)Department of Computing,

the Hong Kong Polytechnic University, Hong Kong. (3)Shenzhen Engineering

Laboratory of Performance Robots at Digital Stage, Shenzhen Graduate School,

Harbin Institute of Technology, Shenzhen, China. (4)School of Engineering and

Applied Science, Aston University, UK.

Protein-DNA interactions are involved in many fundamental biological processes

essential for cellular function. Most of the existing computational approaches

employed only the sequence context of the target residue for its prediction. In

the present study, for each target residue, we applied both the spatial context

and the sequence context to construct the feature space. Subsequently, Latent

Semantic Analysis (LSA) was applied to remove the redundancies in the feature

space. Finally, a predictor (PDNAsite) was developed through the integration of

the support vector machines (SVM) classifier and ensemble learning. Results on

the PDNA-62 and the PDNA-224 datasets demonstrate that features extracted from

spatial context provide more information than those from sequence context and the

combination of them gives more performance gain. An analysis of the number of

binding sites in the spatial context of the target site indicates that the

interactions between binding sites next to each other are important for

protein-DNA recognition and their binding ability. The comparison between our

proposed PDNAsite method and the existing methods indicate that PDNAsite

outperforms most of the existing methods and is a useful tool for DNA-binding

site identification. A web-server of our predictor

(http://hlt.hitsz.edu.cn:8080/PDNAsite/) is made available for free public

accessible to the biological research community.

DOI: 10.1038/srep27653

PMCID: PMC4901350

PMID: 27282833

242. BMC Bioinformatics. 2016 Jun 7;17(1):231. doi: 10.1186/s12859-016-1110-x.

Accurate prediction of RNA-binding protein residues with two discriminative

structural descriptors.

Sun M(1), Wang X(1), Zou C(1), He Z(1), Liu W(1), Li H(2).

Author information:

(1)State Key Laboratory of Bioreactor Engineering, Shanghai Key Laboratory of New

Drug Design, School of Pharmacy, East China University of Science and Technology,

130 Mei Long Road, Shanghai, 200237, China. (2)State Key Laboratory of Bioreactor

Engineering, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East

China University of Science and Technology, 130 Mei Long Road, Shanghai, 200237,

China. hlli@ecust.edu.cn.

BACKGROUND: RNA-binding proteins participate in many important biological

processes concerning RNA-mediated gene regulation, and several computational

methods have been recently developed to predict the protein-RNA interactions of

RNA-binding proteins. Newly developed discriminative descriptors will help to

improve the prediction accuracy of these prediction methods and provide further

meaningful information for researchers.

RESULTS: In this work, we designed two structural features (residue electrostatic

surface potential and triplet interface propensity) and according to the

statistical and structural analysis of protein-RNA complexes, the two features

were powerful for identifying RNA-binding protein residues. Using these two

features and other excellent structure- and sequence-based features, a random

forest classifier was constructed to predict RNA-binding residues. The area under

the receiver operating characteristic curve (AUC) of five-fold cross-validation

for our method on training set RBP195 was 0.900, and when applied to the test set

RBP68, the prediction accuracy (ACC) was 0.868, and the F-score was 0.631.

CONCLUSIONS: The good prediction performance of our method revealed that the two

newly designed descriptors could be discriminative for inferring protein residues

interacting with RNAs. To facilitate the use of our method, a web-server called

RNAProSite, which implements the proposed method, was constructed and is freely

available at http://lilab.ecust.edu.cn/NABind .

DOI: 10.1186/s12859-016-1110-x

PMCID: PMC4897909

PMID: 27266516

243. Database (Oxford). 2016 Jun 7;2016. pii: baw095. doi: 10.1093/database/baw095.

Print 2016.

SinEx DB: a database for single exon coding sequences in mammalian genomes.

Jorquera R(1), Ortiz R(1), Ossandon F(1), Cárdenas JP(1), Sepúlveda R(1),

González C(1), Holmes DS(2).

Author information:

(1)Center for Bioinformatics and Genome Biology, Fundacion Ciencia & Vida and

Facultad de Ciencias Biologicas, Universidad Andres Bello, Avda Zañartu 1482,

Santiago, Chile. (2)Center for Bioinformatics and Genome Biology, Fundacion

Ciencia & Vida and Facultad de Ciencias Biologicas, Universidad Andres Bello,

Avda Zañartu 1482, Santiago, Chile dsholmes2000@yahoo.com.

Eukaryotic genes are typically interrupted by intragenic, noncoding sequences

termed introns. However, some genes lack introns in their coding sequence (CDS)

and are generally known as 'single exon genes' (SEGs). In this work, a SEG is

defined as a nuclear, protein-coding gene that lacks introns in its CDS. Whereas,

many public databases of Eukaryotic multi-exon genes are available, there are

only two specialized databases for SEGs. The present work addresses the need for

a more extensive and diverse database by creating SinEx DB, a publicly available,

searchable database of predicted SEGs from 10 completely sequenced mammalian

genomes including human. SinEx DB houses the DNA and protein sequence information

of these SEGs and includes their functional predictions (KOG) and the relative

distribution of these functions within species. The information is stored in a

relational database built with My SQL Server 5.1.33 and the complete dataset of

SEG sequences and their functional predictions are available for downloading.

SinEx DB can be interrogated by: (i) a browsable phylogenetic schema, (ii)

carrying out BLAST searches to the in-house SinEx DB of SEGs and (iii) via an

advanced search mode in which the database can be searched by key words and any

combination of searches by species and predicted functions. SinEx DB provides a

rich source of information for advancing our understanding of the evolution and

function of SEGs.Database URL: www.sinex.cl.

© The Author(s) 2016. Published by Oxford University Press.

DOI: 10.1093/database/baw095

PMCID: PMC4897596

PMID: 27278816

244. J Biomed Semantics. 2016 Jun 7;7:35. doi: 10.1186/s13326-016-0072-2.

PhenoImageShare: an image annotation and query infrastructure.

Adebayo S(1), McLeod K(2), Tudose I(3), Osumi-Sutherland D(4), Burdett T(4),

Baldock R(1), Burger A(2), Parkinson H(4).

Author information:

(1)MRC Human Genetics Unit, IGMM, University of Edinburgh, Crewe Road, Edinburgh,

UK. (2)Department of Computer Science, Heriot-Watt University, Edinburgh, UK.

(3)European Bioinformatics Institute (EMBL-EBI), European Molecular Biology

Laboratory, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK.

tudose@ebi.ac.uk. (4)European Bioinformatics Institute (EMBL-EBI), European

Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, CB10 1SD,

UK.

BACKGROUND: High throughput imaging is now available to many groups and it is

possible to generate a large quantity of high quality images quickly. Managing

this data, consistently annotating it, or making it available to the community

are all challenges that come with these methods.

RESULTS: PhenoImageShare provides an ontology-enabled lightweight image data

query, annotation service and a single point of access backed by a Solr server

for programmatic access to an integrated image collection enabling improved

community access. PhenoImageShare also provides an easy to use online image

annotation tool with functionality to draw regions of interest on images and to

annotate them with terms from an autosuggest-enabled ontology-lookup widget. The

provenance of each image, and annotation, is kept and links to original resources

are provided. The semantic and intuitive search interface is species and imaging

technology neutral. PhenoImageShare now provides access to annotation for over

100,000 images for 2 species.

CONCLUSION: The PhenoImageShare platform provides underlying infrastructure for

both programmatic access and user-facing tools for biologists enabling the query

and annotation of federated images. PhenoImageShare is accessible online at

http://www.phenoimageshare.org .

DOI: 10.1186/s13326-016-0072-2

PMCID: PMC4896029

PMID: 27267125

245. J Theor Biol. 2016 Jun 7;398:96-102. doi: 10.1016/j.jtbi.2016.03.030. Epub 2016

Mar 26.

Identification of S-glutathionylation sites in species-specific proteins by

incorporating five sequence-derived features into the general pseudo-amino acid

composition.

Zhao X(1), Ning Q(2), Ai M(2), Chai H(2), Yang G(2).

Author information:

(1)School of Computer Science and Information Technology, Northeast Normal

University, Changchun 130117, China. Electronic address: zhaoxw303@nenu.edu.cn.

(2)School of Computer Science and Information Technology, Northeast Normal

University, Changchun 130117, China.

As a selective and reversible protein post-translational modification,

S-glutathionylation generates mixed disulfides between glutathione (GSH) and

cysteine residues, and plays an important role in regulating protein activity,

stability, and redox regulation. To fully understand S-glutathionylation

mechanisms, identification of substrates and specific S-Glutathionylated sites is

crucial. Experimental identification of S-glutathionylated sites is

labor-intensive and time consuming, so establishing an effective computational

method is much desirable due to their convenient and fast speed. Therefore, in

this study, a new bioinformatics tool named SSGlu (Species-Specific

identification of Protein S-glutathionylation Sites) was developed to identify

species-specific protein S-glutathionylated sites, utilizing support vector

machines that combine multiple sequence-derived features with a two-step feature

selection. By 5-fold cross validation, the performance of SSGlu was measured with

an AUC of 0.8105 and 0.8041 for Homo sapiens and Mus musculus, respectively.

Additionally, SSGlu was compared with the existing methods, and the higher MCC

and AUC of SSGlu demonstrated that SSGlu was very promising to predict

S-glutathionylated sites. Furthermore, a site-specific analysis showed that

S-glutathionylation intimately correlated with the features derived from its

surrounding sites. The conclusions derived from this study might help to

understand more of the S-glutathionylation mechanism and guide the related

experimental validation. For public access, SSGlu is freely accessible at

http://59.73.198.144:8080/SSGlu/.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2016.03.030

PMID: 27025952

246. J Theor Biol. 2016 Jun 7;398:122-9. doi: 10.1016/j.jtbi.2016.03.025. Epub 2016

Mar 25.

Identification of WD40 repeats by secondary structure-aided profile-profile

alignment.

Wang C(1), Dong X(2), Han L(3), Su XD(4), Zhang Z(5), Li J(6), Song J(7).

Author information:

(1)State Key Laboratory of Agrobiotechnology, College of Biological Sciences,

China Agricultural University, Beijing 100193, China; Department of Plant

Biology, Carnegie Institution for Science, Stanford, CA 94305, USA. Electronic

address: cwang@carnegiescience.edu. (2)State Key Laboratory of Agrobiotechnology,

College of Biological Sciences, China Agricultural University, Beijing 100193,

China. Electronic address: yndxb@126.com. (3)Center for Cancer Molecular

Diagnosis, Tianjin Medical University Cancer Institute and Hospital, National

Clinical Research Center for Cancer, Tianjin, China. Electronic address:

hanlei\_45@126.com. (4)State Key Laboratory of Protein and Plant Gene Research and

Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, Peking

University, Beijing 100871, China. Electronic address: xdsu@pku.edu.cn. (5)State

Key Laboratory of Agrobiotechnology, College of Biological Sciences, China

Agricultural University, Beijing 100193, China. Electronic address:

zidingzhang@cau.edu.cn. (6)Advanced Analytics Institute and Centre for Health

Technologies, University of Technology Sydney, 81 Broadway, Sydney, NSW 2007,

Australia. Electronic address: Jinyan.Li@uts.edu.au. (7)National Engineering

Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial

Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of

Sciences, Tianjin 300308, China; Infection and Immunity Program, Biomedicine

Discovery Institute and Department of Biochemistry and Molecular Biology, Faculty

of Medicine, Monash University, Melbourne, VIC 3800, Australia; Monash Centre for

Data Science, Faculty of Information Technology, Monash University, Melbourne,

VIC 3800, Australia. Electronic address: jiangning.song@monash.edu.

A WD40 protein typically contains four or more repeats of ~40 residues ended with

the Trp-Asp dipeptide, which folds into β-propellers with four β strands in each

repeat. They often function as scaffolds for protein-protein interactions and are

involved in numerous fundamental biological processes. Despite their important

functional role, the "velcro" closure of WD40 propellers and the diversity of

WD40 repeats make their identification a difficult task. Here we develop a new

WD40 Repeat Recognition method (WDRR), which uses predicted secondary structure

information to generate candidate repeat segments, and further employs a

profile-profile alignment to identify the correct WD40 repeats from candidate

segments. In particular, we design a novel alignment scoring function that

combines dot product and BLOSUM62, thereby achieving a great balance of

sensitivity and accuracy. Taking advantage of these strategies, WDRR could

effectively reduce the false positive rate and accurately identify more remote

homologous WD40 repeats with precise repeat boundaries. We further use WDRR to

re-annotate the Pfam families in the β-propeller clan (CL0186) and identify a

number of WD40 repeat proteins with high confidence across nine model organisms.

The WDRR web server and the datasets are available at

http://protein.cau.edu.cn/wdrr/.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2016.03.025

PMID: 27021623

247. J Theor Biol. 2016 Jun 7;398:32-42. doi: 10.1016/j.jtbi.2016.03.013. Epub 2016

Mar 19.

Mem-ADSVM: A two-layer multi-label predictor for identifying multi-functional

types of membrane proteins.

Wan S(1), Mak MW(2), Kung SY(3).

Author information:

(1)Department of Electronic and Information Engineering, The Hong Kong

Polytechnic University, Hong Kong SAR, China. Electronic address:

shibiao.wan@connect.polyu.hk. (2)Department of Electronic and Information

Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China.

Electronic address: enmwmak@polyu.edu.hk. (3)Department of Electrical

Engineering, Princeton University, NJ, USA. Electronic address:

kung@princeton.edu.

Identifying membrane proteins and their multi-functional types is an

indispensable yet challenging topic in proteomics and bioinformatics. However,

most of the existing membrane-protein predictors have the following problems: (1)

they do not predict whether a given protein is a membrane protein or not; (2)

they are limited to predicting membrane proteins with single-label functional

types but ignore those with multi-functional types; and (3) there is still much

room for improvement for their performance. To address these problems, this paper

proposes a two-layer multi-label predictor, namely Mem-ADSVM, which can identify

membrane proteins (Layer I) and their multi-functional types (Layer II).

Specifically, given a query protein, its associated gene ontology (GO)

information is retrieved by searching a compact GO-term database with its

homologous accession number. Subsequently, the GO information is classified by a

binary support vector machine (SVM) classifier to determine whether it is a

membrane protein or not. If yes, it will be further classified by a multi-label

multi-class SVM classifier equipped with an adaptive-decision (AD) scheme to

determine to which functional type(s) it belongs. Experimental results show that

Mem-ADSVM significantly outperforms state-of-the-art predictors in terms of

identifying both membrane proteins and their multi-functional types. This paper

also suggests that the two-layer prediction architecture is better than the

one-layer for prediction performance. For reader׳s convenience, the Mem-ADSVM

server is available online at http://bioinfo.eie.polyu.edu.hk/MemADSVMServer/.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2016.03.013

PMID: 27000774

248. Oncotarget. 2016 Jun 7;7(23):34558-70. doi: 10.18632/oncotarget.9148.

iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and

incorporating sequence coupled effects into general PseAAC.

Jia J(1,)(2), Liu Z(3), Xiao X(1,)(2), Liu B(1), Chou KC(2,)(4).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403

China. (2)Gordon Life Science Institute, Boston, MA 02478, USA. (3)School of

Computer Science and Engineering, Nanjing University of Science and Technology,

Nanjing, 210094, China. (4)Center of Excellence in Genomic Medicine Research

(CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia.

Carbonylation is a posttranslational modification (PTM or PTLM), where a carbonyl

group is added to lysine (K), proline (P), arginine (R), and threonine (T)

residue of a protein molecule. Carbonylation plays an important role in

orchestrating various biological processes but it is also associated with many

diseases such as diabetes, chronic lung disease, Parkinson's disease, Alzheimer's

disease, chronic renal failure, and sepsis. Therefore, from the angles of both

basic research and drug development, we are facing a challenging problem: for an

uncharacterized protein sequence containing many residues of K, P, R, or T, which

ones can be carbonylated, and which ones cannot? To address this problem, we have

developed a predictor called iCar-PseCp by incorporating the sequence-coupled

information into the general pseudo amino acid composition, and balancing out

skewed training dataset by Monte Carlo sampling to expand positive subset.

Rigorous target cross-validations on a same set of carbonylation-known proteins

indicated that the new predictor remarkably outperformed its existing

counterparts. For the convenience of most experimental scientists, a

user-friendly web-server for iCar-PseCp has been established at

http://www.jci-bioinfo.cn/iCar-PseCp, by which users can easily obtain their

desired results without the need to go through the complicated mathematical

equations involved. It has not escaped our notice that the formulation and

approach presented here can also be used to analyze many other problems in

computational proteomics.

DOI: 10.18632/oncotarget.9148

PMCID: PMC5085176

PMID: 27153555

Conflict of interest statement: The authors declare no conflicts of interest.

249. Oncotarget. 2016 Jun 7;7(23):34180-9. doi: 10.18632/oncotarget.9057.

iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating

dinucleotide position-specific propensity into general pseudo nucleotide

composition.

Xiao X(1,)(2,)(3), Ye HX(1), Liu Z(4), Jia JH(1), Chou KC(5,)(3).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333403,

China. (2)Information School, ZheJiang Textile and Fashion College, NingBo,

315211, China. (3)Gordon Life Science Institute, Boston, Massachusetts, 02478,

USA. (4)School of Computer Science and Engineering, Nanjing University of Science

and Technology, Nanjing, 210094, China. (5)Center of Excellence in Genomic

Medicine Research (CEGMR), King Abdulaziz University, Jeddah, 21589, Saudi

Arabia.

DNA replication, occurring in all living organisms and being the basis for

biological inheritance, is the process of producing two identical replicas from

one original DNA molecule. To in-depth understand such an important biological

process and use it for developing new strategy against genetics diseases, the

knowledge of duplication origin sites in DNA is indispensible. With the explosive

growth of DNA sequences emerging in the postgenomic age, it is highly desired to

develop high throughput tools to identify these regions purely based on the

sequence information alone. In this paper, by incorporating the dinucleotide

position-specific propensity information into the general pseudo nucleotide

composition and using the random forest classifier, a new predictor called

iROS-gPseKNC was proposed. Rigorously cross-validations have indicated that the

proposed predictor is significantly better than the best existing method in

sensitivity, specificity, overall accuracy, and stability. Furthermore, a

user-friendly web-server for iROS-gPseKNC has been established at

http://www.jci-bioinfo.cn/iROS-gPseKNC, by which users can easily get their

desired results without the need to bother the complicated mathematics, which

were presented just for the integrity of the methodology itself.

DOI: 10.18632/oncotarget.9057

PMCID: PMC5085147

PMID: 27147572

Conflict of interest statement: The authors declare no conflicts of interest.

250. Sci Rep. 2016 Jun 6;6:27436. doi: 10.1038/srep27436.

IRESPred: Web Server for Prediction of Cellular and Viral Internal Ribosome Entry

Site (IRES).

Kolekar P(1), Pataskar A(1), Kulkarni-Kale U(1), Pal J(2), Kulkarni A(1).

Author information:

(1)Bioinformatics Centre, Savitribai Phule Pune University (Formerly University

of Pune), Pune, Maharashtra, 411 007, India. (2)Department of Biotechnology,

Savitribai Phule Pune University (Formerly University of Pune), Pune,

Maharashtra, 411 007, India.

Cellular mRNAs are predominantly translated in a cap-dependent manner. However,

some viral and a subset of cellular mRNAs initiate their translation in a

cap-independent manner. This requires presence of a structured RNA element, known

as, Internal Ribosome Entry Site (IRES) in their 5' untranslated regions (UTRs).

Experimental demonstration of IRES in UTR remains a challenging task.

Computational prediction of IRES merely based on sequence and structure

conservation is also difficult, particularly for cellular IRES. A web server,

IRESPred is developed for prediction of both viral and cellular IRES using

Support Vector Machine (SVM). The predictive model was built using 35 features

that are based on sequence and structural properties of UTRs and the

probabilities of interactions between UTR and small subunit ribosomal proteins

(SSRPs). The model was found to have 75.51% accuracy, 75.75% sensitivity, 75.25%

specificity, 75.75% precision and Matthews Correlation Coefficient (MCC) of 0.51

in blind testing. IRESPred was found to perform better than the only available

viral IRES prediction server, VIPS. The IRESPred server is freely available at

http://bioinfo.net.in/IRESPred/.

DOI: 10.1038/srep27436

PMCID: PMC4893748

PMID: 27264539

251. J Exp Clin Cancer Res. 2016 Jun 4;35(1):88. doi: 10.1186/s13046-016-0363-6.

RNF8 promotes epithelial-mesenchymal transition of breast cancer cells.

Kuang J(1), Li L(2), Guo L(3), Su Y(4), Wang Y(1), Xu Y(1), Wang X(5), Meng S(1),

Lei L(6), Xu L(6), Shao G(7).

Author information:

(1)Department of Cell Biology, School of Basic Medical Sciences, Peking

University Health Science Center, Beijing, 100191, China. (2)Department of Cell

Biology, School of Basic Medical Sciences, Peking University Health Science

Center, Beijing, 100191, China. lily@bjmu.edu.cn. (3)Department of Pathology,

School of Basic Medical Sciences, Peking University Health Science Center,

Beijing, 100191, China. (4)The Irma H. Russo, MD Breast Cancer Research

Laboratory, Fox Chase Cancer Center-Temple University Health System,

Philadelphia, PA, 19111, USA. (5)Department of Breast Surgery, the First Hospital

of Jilin University, Changchun, 130021, China. (6)Lab of Molecular Imaging,

Health Science Analysis Center, Peking University, Beijing, 100191, China.

(7)Department of Cell Biology, School of Basic Medical Sciences, Peking

University Health Science Center, Beijing, 100191, China. gzshao@bjmu.edu.cn.

BACKGROUND: Epithelial-mesenchymal transition (EMT) is a crucial step for solid

tumor progression and plays an important role in cancer invasion and metastasis.

RNF8 is an ubiquitin E3 ligase with RING domain, and plays essential roles in DNA

damage response and cell cycle regulation. However the role of RNF8 in the

pathogenesis of breast cancer is still unclear.

METHODS: The expression of RNF8 was examined in different types of breast cell

lines by Western Blotting. EMT associated markers were examined by

Immunofluorescence and Western Blotting in MCF-7 when RNF8 was ectopically

overexpressed, or in MDA-MB-231 when RNF8 was depleted. Transwell and wound

healing assays were performed to assess the effect of RNF8 on cell mobility. The

xenograft model was done with nude mice to investigate the role of RNF8 in tumor

metastasis in vivo. Breast tissue arrays were used to examine the expression of

RNF8 by immunohistochemistry. Kaplan-Meier survival analysis for the relationship

between survival time and RNF8 signature in breast cancer was done with an online

tool ( http://kmplot.com/analysis/ ).

RESULTS: RNF8 is overexpressed in highly metastatic breast cancer cell lines.

Overexpression of RNF8 in MCF-7 significantly promoted EMT phenotypes and

facilitated cell migration. On the contrary, silencing of RNF8 in MDA-MB-231

induced MET phenotypes and inhibited cell migration. Furthermore, we proved that

these metastatic behavior promoting effects of RNF8 in breast cancer was

associated with the inactivation of GSK-3β and activation of β-catenin signaling.

With nude mice xenograft model, we found that shRNA mediated-downregulation of

RNF8 reduced tumor metastasis in vivo. In addition, we found that RNF8 expression

was higher in malignant breast cancer than that of the paired normal breast

tissues, and was positively correlated with lymph node metastases and poor

survival time.

CONCLUSIONS: RNF8 induces EMT in the breast cancer cells and promotes breast

cancer metastasis, suggesting that RNF8 could be used as a potential therapeutic

target for the prevention and treatment of breast cancer.

DOI: 10.1186/s13046-016-0363-6

PMCID: PMC4893263

PMID: 27259701

252. Nucleic Acids Res. 2016 Jun 2;44(10):e91. doi: 10.1093/nar/gkw104. Epub 2016 Feb

20.

SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on

sequence-derived features.

Zhou Y(1), Zeng P(2), Li YH(2), Zhang Z(3), Cui Q(4).

Author information:

(1)Department of Biomedical Informatics, School of Basic Medical Sciences, Peking

University, Beijing 100191, China MOE Key Lab of Molecular Cardiovascular

Sciences, Peking University, Beijing 100191, China Center for Noncoding RNA

Medicine, Peking University Health Science Center, Beijing 100191, China

soontide6825@163.com. (2)Department of Biomedical Informatics, School of Basic

Medical Sciences, Peking University, Beijing 100191, China MOE Key Lab of

Molecular Cardiovascular Sciences, Peking University, Beijing 100191, China

Center for Noncoding RNA Medicine, Peking University Health Science Center,

Beijing 100191, China. (3)State Key Laboratory of Agrobiotechnology, College of

Biological Sciences, China Agricultural University, Beijing 100193, China.

(4)Department of Biomedical Informatics, School of Basic Medical Sciences, Peking

University, Beijing 100191, China MOE Key Lab of Molecular Cardiovascular

Sciences, Peking University, Beijing 100191, China Center for Noncoding RNA

Medicine, Peking University Health Science Center, Beijing 100191, China

cuiqinghua@hsc.pku.edu.cn.

N(6)-methyladenosine (m(6)A) is a prevalent RNA methylation modification involved

in the regulation of degradation, subcellular localization, splicing and local

conformation changes of RNA transcripts. High-throughput experiments have

demonstrated that only a small fraction of the m(6)A consensus motifs in

mammalian transcriptomes are modified. Therefore, accurate identification of RNA

m(6)A sites becomes emergently important. For the above purpose, here a

computational predictor of mammalian m(6)A site named SRAMP is established. To

depict the sequence context around m(6)A sites, SRAMP combines three random

forest classifiers that exploit the positional nucleotide sequence pattern, the

K-nearest neighbor information and the position-independent nucleotide pair

spectrum features, respectively. SRAMP uses either genomic sequences or cDNA

sequences as its input. With either kind of input sequence, SRAMP achieves

competitive performance in both cross-validation tests and rigorous independent

benchmarking tests. Analyses of the informative features and overrepresented

rules extracted from the random forest classifiers demonstrate that nucleotide

usage preferences at the distal positions, in addition to those at the proximal

positions, contribute to the classification. As a public prediction server, SRAMP

is freely available at http://www.cuilab.cn/sramp/.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkw104

PMCID: PMC4889921

PMID: 26896799

253. Algorithms Mol Biol. 2016 Jun 1;11:16. doi: 10.1186/s13015-016-0078-4.

eCollection 2016.

Identification of donor splice sites using support vector machine: a

computational approach based on positional, compositional and dependency

features.

Meher PK(1), Sahu TK(2), Rao AR(2), Wahi SD(1).

Author information:

(1)Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research

Institute, New Delhi, 110012 India. (2)Centre for Agricultural Bioinformatics,

ICAR-Indian Agricultural Statistics Research Institute, New Delhi, 110012 India.

BACKGROUND: Identification of splice sites is essential for annotation of genes.

Though existing approaches have achieved an acceptable level of accuracy, still

there is a need for further improvement. Besides, most of the approaches are

species-specific and hence it is required to develop approaches compatible across

species.

RESULTS: Each splice site sequence was transformed into a numeric vector of

length 49, out of which four were positional, four were dependency and 41 were

compositional features. Using the transformed vectors as input, prediction was

made through support vector machine. Using balanced training set, the proposed

approach achieved area under ROC curve (AUC-ROC) of 96.05, 96.96, 96.95, 96.24 %

and area under PR curve (AUC-PR) of 97.64, 97.89, 97.91, 97.90 %, while tested on

human, cattle, fish and worm datasets respectively. On the other hand, AUC-ROC of

97.21, 97.45, 97.41, 98.06 % and AUC-PR of 93.24, 93.34, 93.38, 92.29 % were

obtained, while imbalanced training datasets were used. The proposed approach was

found comparable with state-of-art splice site prediction approaches, while

compared using the bench mark NN269 dataset and other datasets.

CONCLUSIONS: The proposed approach achieved consistent accuracy across different

species as well as found comparable with the existing approaches. Thus, we

believe that the proposed approach can be used as a complementary method to the

existing methods for the prediction of splice sites. A web server named as

'HSplice' has also been developed based on the proposed approach for easy

prediction of 5' splice sites by the users and is freely available at

http://cabgrid.res.in:8080/HSplice.

DOI: 10.1186/s13015-016-0078-4

PMCID: PMC4888255

PMID: 27252772

254. Interdiscip Sci. 2016 Jun;8(2):186-91. doi: 10.1007/s12539-015-0124-9. Epub 2015

Sep 7.

Identifying Antioxidant Proteins by Using Optimal Dipeptide Compositions.

Feng P(1), Chen W(2), Lin H(3).

Author information:

(1)School of Public Health, North China University of Science and Technology,

Tangshan, 063000, China. (2)Department of Physics, School of Sciences, Center for

Genomics and Computational Biology, North China University of Science and

Technology, Tangshan, 063000, China. greatchen@heuu.edu.cn. (3)Key Laboratory for

Neuro Information of Ministry of Education, Center of Bioinformatics, School of

Life Science and Technology, University of Electronic Science and Technology of

China, Chengdu, 610054, China. hlin@uestc.edu.cn.

Antioxidant proteins are a kind of molecules that can terminate cellular and DNA

damages caused by free radical intermediates. The use of antioxidant proteins for

prevention of diseases has been intensively studied in recent years. Thus,

accurate identification of antioxidant proteins is essential for understanding

their roles in pharmacology. In this study, a support vector machine-based

predictor called AodPred was developed for identifying antioxidant proteins. In

this predictor, the sequence was formulated by using the optimal 3-gap dipeptides

obtained by using feature selection method. It was observed by jackknife

cross-validation test that AodPred can achieve an overall accuracy of 74.79 % in

identifying antioxidant proteins. As a user-friendly tool, AodPred is freely

accessible at http://lin.uestc.edu.cn/server/AntioxiPred . To maximize the

convenience of the vast majority of experimental scientists, a step-by-step guide

is provided on how to use the web server to obtain the desired results.

DOI: 10.1007/s12539-015-0124-9

PMID: 26345449

255. PLoS One. 2016 May 26;11(5):e0154567. doi: 10.1371/journal.pone.0154567.

eCollection 2016.

LncRNApred: Classification of Long Non-Coding RNAs and Protein-Coding Transcripts

by the Ensemble Algorithm with a New Hybrid Feature.

Pian C(1), Zhang G(1), Chen Z(1), Chen Y(1), Zhang J(1), Yang T(1), Zhang L(1).

Author information:

(1)Department of Mathematics, College of Science, Nanjing Agricultural

University, Nanjing, Jiangsu, People's Republic of China.

As a novel class of noncoding RNAs, long noncoding RNAs (lncRNAs) have been

verified to be associated with various diseases. As large scale transcripts are

generated every year, it is significant to accurately and quickly identify

lncRNAs from thousands of assembled transcripts. To accurately discover new

lncRNAs, we develop a classification tool of random forest (RF) named LncRNApred

based on a new hybrid feature. This hybrid feature set includes three new

proposed features, which are MaxORF, RMaxORF and SNR. LncRNApred is effective for

classifying lncRNAs and protein coding transcripts accurately and quickly.

Moreover,our RF model only requests the training using data on human coding and

non-coding transcripts. Other species can also be predicted by using LncRNApred.

The result shows that our method is more effective compared with the Coding

Potential Calculate (CPC). The web server of LncRNApred is available for free at

http://mm20132014.wicp.net:57203/LncRNApred/home.jsp.

DOI: 10.1371/journal.pone.0154567

PMCID: PMC4882039

PMID: 27228152

256. PLoS Comput Biol. 2016 May 25;12(5):e1004962. doi: 10.1371/journal.pcbi.1004962.

eCollection 2016.

PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by

Exploiting the Different Characteristics of Variants in Distinct Genomic Regions.

Bendl J(1,)(2,)(3), Musil M(1,)(2), Štourač J(1,)(3), Zendulka J(2), Damborský

J(1,)(3), Brezovský J(1,)(3).

Author information:

(1)Loschmidt Laboratories, Department of Experimental Biology and Research Centre

for Toxic Compounds in the Environment RECETOX, Masaryk University, Brno, Czech

Republic. (2)Department of Information Systems, Faculty of Information

Technology, Brno University of Technology, Brno, Czech Republic. (3)International

Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech

Republic.

An important message taken from human genome sequencing projects is that the

human population exhibits approximately 99.9% genetic similarity. Variations in

the remaining parts of the genome determine our identity, trace our history and

reveal our heritage. The precise delineation of phenotypically causal variants

plays a key role in providing accurate personalized diagnosis, prognosis, and

treatment of inherited diseases. Several computational methods for achieving such

delineation have been reported recently. However, their ability to pinpoint

potentially deleterious variants is limited by the fact that their mechanisms of

prediction do not account for the existence of different categories of variants.

Consequently, their output is biased towards the variant categories that are most

strongly represented in the variant databases. Moreover, most such methods

provide numeric scores but not binary predictions of the deleteriousness of

variants or confidence scores that would be more easily understood by users. We

have constructed three datasets covering different types of disease-related

variants, which were divided across five categories: (i) regulatory, (ii)

splicing, (iii) missense, (iv) synonymous, and (v) nonsense variants. These

datasets were used to develop category-optimal decision thresholds and to

evaluate six tools for variant prioritization: CADD, DANN, FATHMM, FitCons,

FunSeq2 and GWAVA. This evaluation revealed some important advantages of the

category-based approach. The results obtained with the five best-performing tools

were then combined into a consensus score. Additional comparative analyses showed

that in the case of missense variations, protein-based predictors perform better

than DNA sequence-based predictors. A user-friendly web interface was developed

that provides easy access to the five tools' predictions, and their consensus

scores, in a user-understandable format tailored to the specific features of

different categories of variations. To enable comprehensive evaluation of

variants, the predictions are complemented with annotations from eight databases.

The web server is freely available to the community at

http://loschmidt.chemi.muni.cz/predictsnp2.

DOI: 10.1371/journal.pcbi.1004962

PMCID: PMC4880439

PMID: 27224906

257. PLoS Comput Biol. 2016 May 25;12(5):e1004925. doi: 10.1371/journal.pcbi.1004925.

eCollection 2016.

PEPIS: A Pipeline for Estimating Epistatic Effects in Quantitative Trait Locus

Mapping and Genome-Wide Association Studies.

Zhang W(1), Dai X(1), Wang Q(2), Xu S(3), Zhao PX(1).

Author information:

(1)Plant Biology Division, Samuel Roberts Noble Foundation, Ardmore, Oklahoma,

United States of America. (2)School of Agriculture and Biology, Shanghai Jiao

Tong University, Shanghai, People's Republic of China. (3)Department of Botany

and Plant Sciences, University of California, Riverside, Riverside, California,

United States of America.

The term epistasis refers to interactions between multiple genetic loci. Genetic

epistasis is important in regulating biological function and is considered to

explain part of the 'missing heritability,' which involves marginal genetic

effects that cannot be accounted for in genome-wide association studies. Thus,

the study of epistasis is of great interest to geneticists. However, estimating

epistatic effects for quantitative traits is challenging due to the large number

of interaction effects that must be estimated, thus significantly increasing

computing demands. Here, we present a new web server-based tool, the Pipeline for

estimating EPIStatic genetic effects (PEPIS), for analyzing polygenic epistatic

effects. The PEPIS software package is based on a new linear mixed model that has

been used to predict the performance of hybrid rice. The PEPIS includes two main

sub-pipelines: the first for kinship matrix calculation, and the second for

polygenic component analyses and genome scanning for main and epistatic effects.

To accommodate the demand for high-performance computation, the PEPIS utilizes

C/C++ for mathematical matrix computing. In addition, the modules for kinship

matrix calculations and main and epistatic-effect genome scanning employ parallel

computing technology that effectively utilizes multiple computer nodes across our

networked cluster, thus significantly improving the computational speed. For

example, when analyzing the same immortalized F2 rice population genotypic data

examined in a previous study, the PEPIS returned identical results at each

analysis step with the original prototype R code, but the computational time was

reduced from more than one month to about five minutes. These advances will help

overcome the bottleneck frequently encountered in genome wide epistatic genetic

effect analysis and enable accommodation of the high computational demand. The

PEPIS is publically available at http://bioinfo.noble.org/PolyGenic\_QTL/.

DOI: 10.1371/journal.pcbi.1004925

PMCID: PMC4880203

PMID: 27224861

258. Sci Rep. 2016 May 23;6:26447. doi: 10.1038/srep26447.

MetaTrans: an open-source pipeline for metatranscriptomics.

Martinez X(1), Pozuelo M(1), Pascal V(1), Campos D(1), Gut I(2,)(3), Gut

M(2,)(3), Azpiroz F(1,)(4), Guarner F(1,)(4), Manichanh C(1,)(4).

Author information:

(1)Digestive Research Unit, Vall d'Hebron Research Institute, Barcelona, 08035,

Spain. (2)CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona, 08028, Spain.

(3)Universitat Pompeu Fabra (UPF), 08002, Barcelona, Spain. (4)CIBERehd,

Instituto de Salud Carlos III, Madrid, 28029, Spain.

To date, meta-omic approaches use high-throughput sequencing technologies, which

produce a huge amount of data, thus challenging modern computers. Here we present

MetaTrans, an efficient open-source pipeline to analyze the structure and

functions of active microbial communities using the power of multi-threading

computers. The pipeline is designed to perform two types of RNA-Seq analyses:

taxonomic and gene expression. It performs quality-control assessment, rRNA

removal, maps reads against functional databases and also handles differential

gene expression analysis. Its efficacy was validated by analyzing data from

synthetic mock communities, data from a previous study and data generated from

twelve human fecal samples. Compared to an existing web application server,

MetaTrans shows more efficiency in terms of runtime (around 2 hours per million

of transcripts) and presents adapted tools to compare gene expression levels. It

has been tested with a human gut microbiome database but also proposes an option

to use a general database in order to analyze other ecosystems. For the

installation and use of the pipeline, we provide a detailed guide at the

following website (www.metatrans.org).

DOI: 10.1038/srep26447

PMCID: PMC4876386

PMID: 27211518

259. Genet Mol Res. 2016 May 20;15(2). doi: 10.4238/gmr.15027618.

Identification of Ca(2+)-binding residues of a protein from its primary sequence.

Jiang Z(1), Hu XZ(1), Geriletu G(1), Xing HR(1), Cao XY(1).

Author information:

(1)College of Sciences, Inner Mongolia University of Technology, Hohhot, China.

Calcium is one of the most abundant minerals in the human body, playing a

critical role in many cellular activities by interacting with different calcium

ion (Ca(2+))-binding proteins. Therefore, the correct identification of

Ca(2+)-binding residues is essential for protein functional research. In this

study, a new method was developed to predict Ca2+-binding residues from the

primary sequence without using three-dimensional information. Through statistical

analysis, four kinds of feature parameters were extracted from amino acid

sequences: the increment of diversity values of amino acid composition, the

matrix scoring values of position conservation, the autocross covariance of

physicochemical properties, and the center motif. These features served as input

for a support vector machine to predict Ca(2+)-binding residues. This method was

tested on four well-established datasets using a five-fold cross-validation. The

accuracies and Matthews correlation coefficients were 75.9% and 0.53 (dataset 1),

79.2% and 0.58 (dataset 2), 77.4% and 0.55 (dataset 3), and 79.1% and 0.58

(dataset 4). Comparative results show that the developed method outperforms

previous methods. Based on this study, a web server was developed for predicting

Ca(2+)-binding residues from any protein sequence, being publically available at

http://202.207.29.245/.

DOI: 10.4238/gmr.15027618

PMID: 27323050 [Indexed for MEDLINE]

260. Int J Mol Sci. 2016 May 18;17(5). pii: E757. doi: 10.3390/ijms17050757.

RVMAB: Using the Relevance Vector Machine Model Combined with Average Blocks to

Predict the Interactions of Proteins from Protein Sequences.

An JY(1), You ZH(2), Meng FR(3), Xu SJ(4), Wang Y(5).

Author information:

(1)School of Computer Science and Technology, China University of Mining and

Technology, Xuzhou 21116, China. ajy@cumt.edu.cn. (2)School of Computer Science

and Technology, China University of Mining and Technology, Xuzhou 21116, China.

zhuhongyou@cumt.edu.cn. (3)School of Computer Science and Technology, China

University of Mining and Technology, Xuzhou 21116, China. mengfr@cumt.edu.cn.

(4)School of Computer Science and Technology, China University of Mining and

Technology, Xuzhou 21116, China. ajysjm@163.com. (5)School of Computer Science

and Technology, China University of Mining and Technology, Xuzhou 21116, China.

yinwang@cumt.edu.cn.

Protein-Protein Interactions (PPIs) play essential roles in most cellular

processes. Knowledge of PPIs is becoming increasingly more important, which has

prompted the development of technologies that are capable of discovering

large-scale PPIs. Although many high-throughput biological technologies have been

proposed to detect PPIs, there are unavoidable shortcomings, including cost, time

intensity, and inherently high false positive and false negative rates. For the

sake of these reasons, in silico methods are attracting much attention due to

their good performances in predicting PPIs. In this paper, we propose a novel

computational method known as RVM-AB that combines the Relevance Vector Machine

(RVM) model and Average Blocks (AB) to predict PPIs from protein sequences. The

main improvements are the results of representing protein sequences using the AB

feature representation on a Position Specific Scoring Matrix (PSSM), reducing the

influence of noise using a Principal Component Analysis (PCA), and using a

Relevance Vector Machine (RVM) based classifier. We performed five-fold

cross-validation experiments on yeast and Helicobacter pylori datasets, and

achieved very high accuracies of 92.98% and 95.58% respectively, which is

significantly better than previous works. In addition, we also obtained good

prediction accuracies of 88.31%, 89.46%, 91.08%, 91.55%, and 94.81% on other five

independent datasets C. elegans, M. musculus, H. sapiens, H. pylori, and E. coli

for cross-species prediction. To further evaluate the proposed method, we compare

it with the state-of-the-art support vector machine (SVM) classifier on the yeast

dataset. The experimental results demonstrate that our RVM-AB method is obviously

better than the SVM-based method. The promising experimental results show the

efficiency and simplicity of the proposed method, which can be an automatic

decision support tool. To facilitate extensive studies for future proteomics

research, we developed a freely available web server called RVMAB-PPI in

Hypertext Preprocessor (PHP) for predicting PPIs. The web server including source

code and the datasets are available at http://219.219.62.123:8888/ppi\_ab/.

DOI: 10.3390/ijms17050757

PMCID: PMC4881578

PMID: 27213337

261. IEEE/ACM Trans Comput Biol Bioinform. 2016 May 16. [Epub ahead of print]

txCoords: A Novel Web Application for Transcriptomic Peak Re-mapping.

Yan Z, Liu K, Xiang S, Sun Z.

Since the development of new technologies such as RIP-Seq and m6A-seq, peak

calling has become an important step in transcriptomic sequencing data analysis.

However, many of the reported genomic coordinates of transcriptomic peaks are

incorrect owing to negligence of the introns. There is currently a lack of a

convenient tool to address this problem. Here, we present txCoords, a novel and

easy-to-use web application for transcriptomic peak re-mapping. txCoords can be

used to correct the incorrectly reported transcriptomic peaks and retrieve the

true sequences. It also supports visualization of the re-mapped peaks in a

schematic figure or from the UCSC Genome Browser. Our web server is freely

available at http://www.bioinfo.tsinghua.edu.cn/txCoords.

DOI: 10.1109/TCBB.2016.2568178

PMID: 27214906

262. Bioinformatics. 2016 May 15;32(10):1598-600. doi: 10.1093/bioinformatics/btw043.

Epub 2016 Jan 23.

BISQUE: locus- and variant-specific conversion of genomic, transcriptomic and

proteomic database identifiers.

Meyer MJ(1), Geske P(2), Yu H(2).

Author information:

(1)Department of Biological Statistics and Computational Biology, Weill Institute

for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA and

Tri-Institutional Training Program in Computational Biology and Medicine, New

York, NY 10065, USA. (2)Department of Biological Statistics and Computational

Biology, Weill Institute for Cell and Molecular Biology, Cornell University,

Ithaca, NY 14853, USA and.

Biological sequence databases are integral to efforts to characterize and

understand biological molecules and share biological data. However, when

analyzing these data, scientists are often left holding disparate biological

currency-molecular identifiers from different databases. For downstream

applications that require converting the identifiers themselves, there are many

resources available, but analyzing associated loci and variants can be cumbersome

if data is not given in a form amenable to particular analyses. Here we present

BISQUE, a web server and customizable command-line tool for converting molecular

identifiers and their contained loci and variants between different database

conventions. BISQUE uses a graph traversal algorithm to generalize the conversion

process for residues in the human genome, genes, transcripts and proteins,

allowing for conversion across classes of molecules and in all directions through

an intuitive web interface and a URL-based web service.AVAILABILITY AND

IMPLEMENTATION: BISQUE is freely available via the web using any major web

browser (http://bisque.yulab.org/). Source code is available in a public GitHub

repository (https://github.com/hyulab/BISQUE).

CONTACT: haiyuan.yu@cornell.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw043

PMID: 26803163

263. Bioinformatics. 2016 May 15;32(10):1574-6. doi: 10.1093/bioinformatics/btw036.

Epub 2016 Jan 22.

Alloscore: a method for predicting allosteric ligand-protein interactions.

Li S(1), Shen Q(1), Su M(2), Liu X(1), Lu S(1), Chen Z(1), Wang R(2), Zhang J(1).

Author information:

(1)Department of Pathophysiology, Shanghai Jiao-Tong University School of

Medicine, Shanghai 200025, China and. (2)Shanghai Institute of Organic Chemistry,

Chinese Academy of Sciences, Shanghai 200032, China.

Allosteric ligands have increasingly gained attention as potential therapeutic

agents due to their higher target selectivity and lower toxicity compared with

classic orthosteric ligands. Despite the great interest in the development of

allosteric drugs as a new tactic in drug discovery, the understanding of the

ligand-protein interactions underlying allosteric binding represents a key

challenge. Herein, we introduce Alloscore, a web server that predicts the binding

affinities of allosteric ligand-protein interactions. This method exhibits

prominent performance in describing allosteric binding and could be useful in

allosteric virtual screening and the structural optimization of allosteric

agonists/antagonists.AVAILABILITY AND IMPLEMENTATION: The Alloscore server and

tutorials are freely available at http://mdl.shsmu.edu.cn/alloscore

CONTACT: jian.zhang@sjtu.edu.cn

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw036

PMID: 26803160

264. Bioinformatics. 2016 May 15;32(10):1589-91. doi: 10.1093/bioinformatics/btw031.

Epub 2016 Jan 21.

ELASPIC web-server: proteome-wide structure-based prediction of mutation effects

on protein stability and binding affinity.

Witvliet DK(1), Strokach A(2), Giraldo-Forero AF(3), Teyra J(4), Colak R(2), Kim

PM(5).

Author information:

(1)Donnelly Center for Cellular and Biomolecular Research, University of Toronto,

Department of Molecular Genetics, University of Toronto. (2)Donnelly Center for

Cellular and Biomolecular Research, University of Toronto, Department of Computer

Science, University of Toronto, Toronto, ON, Canada and. (3)Department of

Electrical and Electronic Engineering, Universidad Nacional De Colombia,

Manizales, Colombia. (4)Donnelly Center for Cellular and Biomolecular Research,

University of Toronto. (5)Donnelly Center for Cellular and Biomolecular Research,

University of Toronto, Department of Molecular Genetics, University of Toronto,

Department of Computer Science, University of Toronto, Toronto, ON, Canada and.

ELASPIC is a novel ensemble machine-learning approach that predicts the effects

of mutations on protein folding and protein-protein interactions. Here, we

present the ELASPIC webserver, which makes the ELASPIC pipeline available through

a fast and intuitive interface. The webserver can be used to evaluate the effect

of mutations on any protein in the Uniprot database, and allows all predicted

results, including modeled wild-type and mutated structures, to be managed and

viewed online and downloaded if needed. It is backed by a database which contains

improved structural domain definitions, and a list of curated domain-domain

interactions for all known proteins, as well as homology models of domains and

domain-domain interactions for the human proteome. Homology models for proteins

of other organisms are calculated on the fly, and mutations are evaluated within

minutes once the homology model is available.AVAILABILITY AND IMPLEMENTATION: The

ELASPIC webserver is available online at http://elaspic.kimlab.org

CONTACT: pm.kim@utoronto.ca or pi@kimlab.orgSupplementary data: Supplementary

data are available at Bioinformatics online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btw031

PMID: 26801957

265. J Comput Chem. 2016 May 15;37(13):1223-9. doi: 10.1002/jcc.24314. Epub 2016 Feb

2.

Sequence-based prediction of protein-peptide binding sites using support vector

machine.

Taherzadeh G(1), Yang Y(1,)(2), Zhang T(3), Liew AW(1), Zhou Y(1,)(2).

Author information:

(1)School of Information and Communication Technology, Griffith University,

Parklands Drive, Southport, Queensland, 4215, Australia. (2)Institute for

Glycomics, Griffith University, Parklands Dr, Southport, Queensland, 4215,

Australia. (3)Weill Cornell Medical College, 1300 York Avenue, New York, New

York, 10065.

Protein-peptide interactions are essential for all cellular processes including

DNA repair, replication, gene-expression, and metabolism. As most protein-peptide

interactions are uncharacterized, it is cost effective to investigate them

computationally as the first step. All existing approaches for predicting

protein-peptide binding sites, however, are based on protein structures despite

the fact that the structures for most proteins are not yet solved. This article

proposes the first machine-learning method called SPRINT to make Sequence-based

prediction of Protein-peptide Residue-level Interactions. SPRINT yields a robust

and consistent performance for 10-fold cross validations and independent test.

The most important feature is evolution-generated sequence profiles. For the test

set (1056 binding and non-binding residues), it yields a Matthews' Correlation

Coefficient of 0.326 with a sensitivity of 64% and a specificity of 68%. This

sequence-based technique shows comparable or more accurate than structure-based

methods for peptide-binding site prediction. SPRINT is available as an online

server at: http://sparks-lab.org/. © 2016 Wiley Periodicals, Inc.

© 2016 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24314

PMID: 26833816

266. Database (Oxford). 2016 May 12;2016. pii: baw070. doi: 10.1093/database/baw070.

Print 2016.

PepPSy: a web server to prioritize gene products in experimental and biocuration

workflows.

Sallou O(1), Duek PD(2), Darde TA(3), Collin O(1), Lane L(4), Chalmel F(5).

Author information:

(1)Genouest Bioinformatics Platform, IRISA, Campus de Beaulieu, Rennes 35042,

France. (2)CALIPHO Group, SIB Swiss Institute of Bioinformatics, CMU, Michel

Servet 1, Geneva 1211, Switzerland. (3)Genouest Bioinformatics Platform, IRISA,

Campus de Beaulieu, Rennes 35042, France IRSET, Inserm U1085, 9 avenue du

Professeur Léon Bernard, Rennes 35000, France. (4)CALIPHO Group, SIB Swiss

Institute of Bioinformatics, CMU, Michel Servet 1, Geneva 1211, Switzerland

Department of Human Protein Sciences, Faculty of Medicine, University of Geneva,

CMU, Michel Servet 1, Geneva 1211, Switzerland. (5)IRSET, Inserm U1085, 9 avenue

du Professeur Léon Bernard, Rennes 35000, France frederic.chalmel@inserm.fr.

Among the 20 000 human gene products predicted from genome annotation, about 3000

still lack validation at protein level. We developed PepPSy, a user-friendly gene

expression-based prioritization system, to help investigators to determine in

which human tissues they should look for an unseen protein. PepPSy can also be

used by biocurators to revisit the annotation of specific categories of proteins

based on the 'omics' data housed by the system. In this study, it was used to

prioritize 21 dubious protein-coding genes among the 616 annotated in neXtProt

for reannotation. PepPSy is freely available at

http://peppsy.genouest.orgDatabase URL: http://peppsy.genouest.org.

© The Author(s) 2016. Published by Oxford University Press.

DOI: 10.1093/database/baw070

PMCID: PMC4865363

PMID: 27173522 [Indexed for MEDLINE]

267. PLoS One. 2016 May 11;11(5):e0155443. doi: 10.1371/journal.pone.0155443.

eCollection 2016.

MiasDB: A Database of Molecular Interactions Associated with Alternative Splicing

of Human Pre-mRNAs.

Xing Y(1), Zhao X(1), Yu T(2), Liang D(1), Li J(1), Wei G(1), Liu G(1), Cui X(1),

Zhao H(1), Cai L(1).

Author information:

(1)School of Life Science and Technology, Inner Mongolia University of Science

and Technology, Baotou, 014010, China. (2)School of Science, Inner Mongolia

University of Science and Technology, Baotou, 014010, China.

Alternative splicing (AS) is pervasive in human multi-exon genes and is a major

contributor to expansion of the transcriptome and proteome diversity. The

accurate recognition of alternative splice sites is regulated by information

contained in networks of protein-protein and protein-RNA interactions. However,

the mechanisms leading to splice site selection are not fully understood.

Although numerous databases have been built to describe AS, molecular interaction

databases associated with AS have only recently emerged. In this study, we

present a new database, MiasDB, that provides a description of molecular

interactions associated with human AS events. This database covers 938

interactions between human splicing factors, RNA elements, transcription factors,

kinases and modified histones for 173 human AS events. Every entry includes the

interaction partners, interaction type, experimental methods, AS type, tissue

specificity or disease-relevant information, a simple description of the

functionally tested interaction in the AS event and references. The database can

be queried easily using a web server (http://47.88.84.236/Miasdb). We display

some interaction figures for several genes. With this database, users can view

the regulation network describing AS events for 12 given genes.

DOI: 10.1371/journal.pone.0155443

PMCID: PMC4864242

PMID: 27167218

268. PLoS One. 2016 May 5;11(5):e0154786. doi: 10.1371/journal.pone.0154786.

eCollection 2016.

Estimation of Uncertainties in the Global Distance Test (GDT\_TS) for CASP Models.

Li W(1), Schaeffer RD(2), Otwinowski Z(1), Grishin NV(2,)(1).

Author information:

(1)Department of Biochemistry and Department of Biophysics, University of Texas

Southwestern Medical Center, Dallas, Texas, 75390-9050, United States of America.

(2)Howard Hughes Medical Institute, University of Texas Southwestern Medical

Center, Dallas, Texas, 75390-9050, United States of America.

The Critical Assessment of techniques for protein Structure Prediction (or CASP)

is a community-wide blind test experiment to reveal the best accomplishments of

structure modeling. Assessors have been using the Global Distance Test (GDT\_TS)

measure to quantify prediction performance since CASP3 in 1998. However,

identifying significant score differences between close models is difficult

because of the lack of uncertainty estimations for this measure. Here, we

utilized the atomic fluctuations caused by structure flexibility to estimate the

uncertainty of GDT\_TS scores. Structures determined by nuclear magnetic resonance

are deposited as ensembles of alternative conformers that reflect the structural

flexibility, whereas standard X-ray refinement produces the static structure

averaged over time and space for the dynamic ensembles. To recapitulate the

structural heterogeneous ensemble in the crystal lattice, we performed

time-averaged refinement for X-ray datasets to generate structural ensembles for

our GDT\_TS uncertainty analysis. Using those generated ensembles, our study

demonstrates that the time-averaged refinements produced structure ensembles with

better agreement with the experimental datasets than the averaged X-ray

structures with B-factors. The uncertainty of the GDT\_TS scores, quantified by

their standard deviations (SDs), increases for scores lower than 50 and 70, with

maximum SDs of 0.3 and 1.23 for X-ray and NMR structures, respectively. We also

applied our procedure to the high accuracy version of GDT-based score and

produced similar results with slightly higher SDs. To facilitate score

comparisons by the community, we developed a user-friendly web server that

produces structure ensembles for NMR and X-ray structures and is accessible at

http://prodata.swmed.edu/SEnCS. Our work helps to identify the significance of

GDT\_TS score differences, as well as to provide structure ensembles for

estimating SDs of any scores.

DOI: 10.1371/journal.pone.0154786

PMCID: PMC4858170

PMID: 27149620

269. Biochim Biophys Acta. 2016 May;1864(5):435-40. doi: 10.1016/j.bbapap.2016.02.005.

Epub 2016 Feb 5.

GprotPRED: Annotation of Gα, Gβ and Gγ subunits of G-proteins using profile

Hidden Markov Models (pHMMs) and application to proteomes.

Kostiou VD(1), Theodoropoulou MC(1), Hamodrakas SJ(2).

Author information:

(1)Department of Cell Biology and Biophysics, Faculty of Biology, University of

Athens, Athens 157 01, Greece. (2)Department of Cell Biology and Biophysics,

Faculty of Biology, University of Athens, Athens 157 01, Greece. Electronic

address: shamodr@biol.uoa.gr.

Heterotrimeric G-proteins form a major protein family, which participates in

signal transduction. They are composed of three subunits, Gα, Gβ and Gγ. The Gα

subunit is further divided in four distinct families Gs, Gi/o, Gq/11 and G12/13.

The goal of this work was to detect and classify members of the four distinct

families, plus the Gβ and the Gγ subunits of G-proteins from sequence alone. To

achieve this purpose, six specific profile Hidden Markov Models (pHMMs) were

built and checked for their credibility. These models were then applied to ten

(10) proteomes and were able to identify all known G-protein and classify them

into the distinct families. In a separate case study, the models were applied to

twenty seven (27) arthropod proteomes and were able to give more credible

classification in proteins with uncertain annotation and in some cases to detect

novel proteins. An online tool, GprotPRED, was developed that uses these six

pHMMs. The sensitivity and specificity for all pHMMs were equal to 100% with the

exception of the Gβ case, where sensitivity equals to 100%, while specificity is

99.993%. In contrast to Pfam's pHMM which detects Gα subunits in general, our

method not only detects Gα subunits but also classifies them into the appropriate

Gα-protein family and thus could become a useful tool for the annotation of

G-proteins in newly discovered proteomes. GprotPRED online tool is publicly

available for non-commercial use at http://bioinformatics.biol.uoa.gr/GprotPRED

and, also, a standalone version of the tool at

https://github.com/vkostiou/GprotPRED.

Copyright © 2016 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.bbapap.2016.02.005

PMID: 26854601 [Indexed for MEDLINE]

270. Bioinformatics. 2016 May 1;32(9):1395-401. doi: 10.1093/bioinformatics/btw013.

Epub 2016 Jan 10.

Collaborative analysis of multi-gigapixel imaging data using Cytomine.

Marée R(1), Rollus L(2), Stévens B(2), Hoyoux R(2), Louppe G(2), Vandaele R(2),

Begon JM(2), Kainz P(3), Geurts P(2), Wehenkel L(2).

Author information:

(1)Systems and Modeling, Department of Electrical Engineering and Computer

Science and GIGA-Research, University of Liège, Liège, Belgium Bioimage Analysis

Unit, Institut Pasteur, Paris, France. (2)Systems and Modeling, Department of

Electrical Engineering and Computer Science and GIGA-Research, University of

Liège, Liège, Belgium. (3)Institute of Biophysics, Medical University of Graz,

Graz, Austria.

MOTIVATION: Collaborative analysis of massive imaging datasets is essential to

enable scientific discoveries.

RESULTS: We developed Cytomine to foster active and distributed collaboration of

multidisciplinary teams for large-scale image-based studies. It uses web

development methodologies and machine learning in order to readily organize,

explore, share and analyze (semantically and quantitatively) multi-gigapixel

imaging data over the internet. We illustrate how it has been used in several

biomedical applications.

AVAILABILITY AND IMPLEMENTATION: Cytomine (http://www.cytomine.be/) is freely

available under an open-source license from http://github.com/cytomine/ A

documentation wiki (http://doc.cytomine.be) and a demo server

(http://demo.cytomine.be) are also available.

CONTACT: info@cytomine.be

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btw013

PMCID: PMC4848407

PMID: 26755625

271. Bioinformatics. 2016 May 1;32(9):1405-7. doi: 10.1093/bioinformatics/btv727. Epub

2016 Jan 5.

VIRALpro: a tool to identify viral capsid and tail sequences.

Galiez C(1), Magnan CN(2), Coste F(1), Baldi P(2).

Author information:

(1)INRIA, Campus De Beaulieu, Rennes Cedex, 35042, France. (2)Department of

Computer Science and Institute for Genomics and Bioinformatics, University of

California, Irvine, Irvine, CA 92697, USA.

MOTIVATION: Not only sequence data continue to outpace annotation information,

but also the problem is further exacerbated when organisms are underrepresented

in the annotation databases. This is the case with non-human-pathogenic viruses

which occur frequently in metagenomic projects. Thus, there is a need for tools

capable of detecting and classifying viral sequences.

RESULTS: We describe VIRALpro a new effective tool for identifying capsid and

tail protein sequences, which are the cornerstones toward viral sequence

annotation and viral genome classification.

AVAILABILITY AND IMPLEMENTATION: The data, software and corresponding web server

are available from http://scratch.proteomics.ics.uci.edu as part of the SCRATCH

suite.

CONTACT: clovis.galiez@inria.fr or pfbaldi@uci.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2016. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv727

PMID: 26733451

272. Hum Mutat. 2016 May;37(5):447-56. doi: 10.1002/humu.22963. Epub 2016 Feb 18.

mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants

in the Structural Proteome.

Meyer MJ(1,)(2,)(3), Lapcevic R(1,)(2), Romero AE(4), Yoon M(1,)(2), Das

J(1,)(2), Beltrán JF(1,)(2), Mort M(5), Stenson PD(5), Cooper DN(5), Paccanaro

A(4), Yu H(1,)(2).

Author information:

(1)Department of Biological Statistics and Computational Biology, Cornell

University, Ithaca, New York, 14853. (2)Weill Institute for Cell and Molecular

Biology, Cornell University, Ithaca, New York, 14853. (3)Tri-Institutional

Training Program in Computational Biology and Medicine, New York, New York,

10065. (4)Department of Computer Science and Centre for Systems and Synthetic

Biology, Royal Holloway, University of London, Egham TW20 0EX, UK. (5)Institute

of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff

CF14 4XN, UK.

A new algorithm and Web server, mutation3D (http://mutation3d.org), proposes

driver genes in cancer by identifying clusters of amino acid substitutions within

tertiary protein structures. We demonstrate the feasibility of using a 3D

clustering approach to implicate proteins in cancer based on explorations of

single proteins using the mutation3D Web interface. On a large scale, we show

that clustering with mutation3D is able to separate functional from nonfunctional

mutations by analyzing a combination of 8,869 known inherited disease mutations

and 2,004 SNPs overlaid together upon the same sets of crystal structures and

homology models. Further, we present a systematic analysis of whole-genome and

whole-exome cancer datasets to demonstrate that mutation3D identifies many known

cancer genes as well as previously underexplored target genes. The mutation3D Web

interface allows users to analyze their own mutation data in a variety of popular

formats and provides seamless access to explore mutation clusters derived from

over 975,000 somatic mutations reported by 6,811 cancer sequencing studies. The

mutation3D Web interface is freely available with all major browsers supported.

© 2016 WILEY PERIODICALS, INC.

DOI: 10.1002/humu.22963

PMCID: PMC4833594 [Available on 2017-05-01]

PMID: 26841357 [Indexed for MEDLINE]

273. J Comput Aided Mol Des. 2016 May;30(5):413-24. doi: 10.1007/s10822-016-9915-2.

Epub 2016 May 11.

TargetNet: a web service for predicting potential drug-target interaction

profiling via multi-target SAR models.

Yao ZJ(1,)(2), Dong J(1), Che YJ(3), Zhu MF(3), Wen M(2), Wang NN(1), Wang S(2),

Lu AP(4), Cao DS(5,)(6).

Author information:

(1)School of Pharmaceutical Sciences, Central South University, Changsha, 410013,

People's Republic of China. (2)College of Chemistry and Chemical Engineering,

Central South University, Changsha, 410083, People's Republic of China. (3)School

of Mathematics and Statistics, Central South University, Changsha, 410083,

People's Republic of China. (4)Institute of Advancing Translational Medicine in

Bone and Joint Diseases, School of Chinese Medicine, Hong Kong Baptist

University, Hong Kong, SAR, People's Republic of China. (5)School of

Pharmaceutical Sciences, Central South University, Changsha, 410013, People's

Republic of China. oriental-cds@163.com. (6)Institute of Advancing Translational

Medicine in Bone and Joint Diseases, School of Chinese Medicine, Hong Kong

Baptist University, Hong Kong, SAR, People's Republic of China.

oriental-cds@163.com.

Drug-target interactions (DTIs) are central to current drug discovery processes

and public health fields. Analyzing the DTI profiling of the drugs helps to infer

drug indications, adverse drug reactions, drug-drug interactions, and drug mode

of actions. Therefore, it is of high importance to reliably and fast predict DTI

profiling of the drugs on a genome-scale level. Here, we develop the TargetNet

server, which can make real-time DTI predictions based only on molecular

structures, following the spirit of multi-target SAR methodology. Naïve Bayes

models together with various molecular fingerprints were employed to construct

prediction models. Ensemble learning from these fingerprints was also provided to

improve the prediction ability. When the user submits a molecule, the server will

predict the activity of the user's molecule across 623 human proteins by the

established high quality SAR model, thus generating a DTI profiling that can be

used as a feature vector of chemicals for wide applications. The 623 SAR models

related to 623 human proteins were strictly evaluated and validated by several

model validation strategies, resulting in the AUC scores of 75-100 %. We applied

the generated DTI profiling to successfully predict potential targets, toxicity

classification, drug-drug interactions, and drug mode of action, which

sufficiently demonstrated the wide application value of the potential DTI

profiling. The TargetNet webserver is designed based on the Django framework in

Python, and is freely accessible at http://targetnet.scbdd.com .

DOI: 10.1007/s10822-016-9915-2

PMID: 27167132

274. Mol Biol Evol. 2016 May;33(5):1205-18. doi: 10.1093/molbev/msw005. Epub 2016 Jan

13.

267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related

Genetic Variation.

Dopazo J(1), Amadoz A(2), Bleda M(3), Garcia-Alonso L(2), Alemán A(3),

García-García F(2), Rodriguez JA(4), Daub JT(4), Muntané G(4), Rueda A(5),

Vela-Boza A(5), López-Domingo FJ(5), Florido JP(5), Arce P(5), Ruiz-Ferrer M(6),

Méndez-Vidal C(7), Arnold TE(8), Spleiss O(9), Alvarez-Tejado M(10), Navarro

A(11), Bhattacharya SS(12), Borrego S(7), Santoyo-López J(5), Antiñolo G(13).

Author information:

(1)Computational Genomics Department, Centro de Investigación Príncipe Felipe

(CIPF), Valencia, Spain Medical Genome Project, Genomics and Bioinformatics

Platform of Andalusia (GBPA), Sevilla, Spain Bioinformatics in Rare Diseases

(BIER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER),

Valencia, Spain Functional Genomics Node, National Institute of Bioinformatics

(INB), Valencia, Spain jdopazo@cipf.es gantinolo@us.es. (2)Computational Genomics

Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain.

(3)Computational Genomics Department, Centro de Investigación Príncipe Felipe

(CIPF), Valencia, Spain Bioinformatics in Rare Diseases (BIER), Centro de

Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Valencia, Spain.

(4)Institut De Biologia Evolutiva, Consejo Superior de Investigaciones

Científicas - Universitat Pompeu Fabra, Barcelona, Spain. (5)Medical Genome

Project, Genomics and Bioinformatics Platform of Andalusia (GBPA), Sevilla,

Spain. (6)Medical Genome Project, Genomics and Bioinformatics Platform of

Andalusia (GBPA), Sevilla, Spain Department of Genetics, Reproduction and Fetal

Medicine, Institute of Biomedicine of Seville, University Hospital Virgen del

Rocío/Consejo Superior de Investigaciones Científicas/University of Seville,

Sevilla, Spain Centro de Investigación Biomédica en Red de Enfermedades Raras

(CIBERER), Sevilla, Spain. (7)Department of Genetics, Reproduction and Fetal

Medicine, Institute of Biomedicine of Seville, University Hospital Virgen del

Rocío/Consejo Superior de Investigaciones Científicas/University of Seville,

Sevilla, Spain Centro de Investigación Biomédica en Red de Enfermedades Raras

(CIBERER), Sevilla, Spain. (8)Research and Development, 454 Life Sciences, a

Roche Company, Branford, CT, USA. (9)Roche Pharma Research and Early Development,

Pharmaceutical Sciences, Roche Innovation Center Basel, Basel, Switzerland.

(10)Roche Diagnostics SL, Sant Cugat del Vallès, Spain. (11)Departament of

Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain

Institució Catalana de Recerca I Estudis Avançats (ICREA), Barcelona Biomedical

Research Park (PRBB), Barcelona, Spain Center for Genomic Regulation (CRG),

Barcelona Biomedical Research Park (PRBB), Barcelona, Spain. (12)Medical Genome

Project, Genomics and Bioinformatics Platform of Andalusia (GBPA), Sevilla, Spain

Andalusian Molecular Biology and Regenerative Medicine Centre (CABIMER), Sevilla,

Spain. (13)Medical Genome Project, Genomics and Bioinformatics Platform of

Andalusia (GBPA), Sevilla, Spain Department of Genetics, Reproduction and Fetal

Medicine, Institute of Biomedicine of Seville, University Hospital Virgen del

Rocío/Consejo Superior de Investigaciones Científicas/University of Seville,

Sevilla, Spain Centro de Investigación Biomédica en Red de Enfermedades Raras

(CIBERER), Sevilla, Spain jdopazo@cipf.es gantinolo@us.es.

Recent results from large-scale genomic projects suggest that allele frequencies,

which are highly relevant for medical purposes, differ considerably across

different populations. The need for a detailed catalog of local variability

motivated the whole-exome sequencing of 267 unrelated individuals, representative

of the healthy Spanish population. Like in other studies, a considerable number

of rare variants were found (almost one-third of the described variants). There

were also relevant differences in allelic frequencies in polymorphic variants,

including ∼10,000 polymorphisms private to the Spanish population. The allelic

frequencies of variants conferring susceptibility to complex diseases (including

cancer, schizophrenia, Alzheimer disease, type 2 diabetes, and other pathologies)

were overall similar to those of other populations. However, the trend is the

opposite for variants linked to Mendelian and rare diseases (including several

retinal degenerative dystrophies and cardiomyopathies) that show marked frequency

differences between populations. Interestingly, a correspondence between

differences in allelic frequencies and disease prevalence was found, highlighting

the relevance of frequency differences in disease risk. These differences are

also observed in variants that disrupt known drug binding sites, suggesting an

important role for local variability in population-specific drug resistances or

adverse effects. We have made the Spanish population variant server web page that

contains population frequency information for the complete list of 170,888

variant positions we found publicly available (http://spv.babelomics.org/), We

show that it if fundamental to determine population-specific variant frequencies

to distinguish real disease associations from population-specific polymorphisms.

© The Author 2016. Published by Oxford University Press on behalf of the Society

for Molecular Biology and Evolution.

DOI: 10.1093/molbev/msw005

PMCID: PMC4839216

PMID: 26764160

275. Sci Rep. 2016 Apr 26;6:24782. doi: 10.1038/srep24782.

A web-based resource for designing therapeutics against Ebola Virus.

Dhanda SK(1), Chaudhary K(1), Gupta S(1), Brahmachari SK(2), Raghava GP(1).

Author information:

(1)Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39A,

Chandigarh, India. (2)CSIR-Institute of Genomics and Integrative Biology, Mathura

Road, New Delhi, India.

In this study, we describe a web-based resource, developed for assisting the

scientific community in designing an effective therapeutics against the Ebola

virus. Firstly, we predicted and identified experimentally validated epitopes in

each of the antigens/proteins of the five known ebolaviruses. Secondly, we

generated all the possible overlapping 9mer peptides from the proteins of

ebolaviruses. Thirdly, conserved peptides across all the five ebolaviruses (four

human pathogenic species) with no identical sequence in the human proteome, based

on 1000 Genomes project, were identified. Finally, we identified peptide or

epitope-based vaccine candidates that could activate both the B- and T-cell arms

of the immune system. In addition, we also identified efficacious siRNAs against

the mRNA transcriptome (absent in human transcriptome) of all the five

ebolaviruses. It was observed that three species can potentially be targeted by a

single siRNA (19mer) and 75 siRNAs can potentially target at least two species. A

web server, EbolaVCR, has been developed that incorporates all the above

information and useful computational tools (http://crdd.osdd.net/oscadd/ebola/).

DOI: 10.1038/srep24782

PMCID: PMC4845023

PMID: 27113850

276. PLoS One. 2016 Apr 22;11(4):e0154237. doi: 10.1371/journal.pone.0154237.

eCollection 2016.

iSulf-Cys: Prediction of S-sulfenylation Sites in Proteins with Physicochemical

Properties of Amino Acids.

Xu Y(1), Ding J(1), Wu LY(2).

Author information:

(1)Department of Information and Computer Science, University of Science and

Technology Beijing, Beijing 100083, China. (2)Institute of Applied Mathematics,

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing

100190, China.

Cysteine S-sulfenylation is an important post-translational modification (PTM) in

proteins, and provides redox regulation of protein functions. Bioinformatics and

structural analyses indicated that S-sulfenylation could impact many biological

and functional categories and had distinct structural features. However, major

limitations for identifying cysteine S-sulfenylation were expensive and

low-throughout. In view of this situation, the establishment of a useful

computational method and the development of an efficient predictor are highly

desired. In this study, a predictor iSulf-Cys which incorporated 14 kinds of

physicochemical properties of amino acids was proposed. With the 10-fold

cross-validation, the value of area under the curve (AUC) was 0.7155 ± 0.0085,

MCC 0.3122 ± 0.0144 on the training dataset for 20 times. iSulf-Cys also showed

satisfying performance in the independent testing dataset with AUC 0.7343 and MCC

0.3315. Features which were constructed from physicochemical properties and

position were carefully analyzed. Meanwhile, a user-friendly web-server for

iSulf-Cys is accessible at http://app.aporc.org/iSulf-Cys/.

DOI: 10.1371/journal.pone.0154237

PMCID: PMC4841585

PMID: 27104833

277. PLoS One. 2016 Apr 20;11(4):e0153771. doi: 10.1371/journal.pone.0153771.

eCollection 2016.

A Web-Based Platform for Designing Vaccines against Existing and Emerging Strains

of Mycobacterium tuberculosis.

Dhanda SK(1), Vir P(1), Singla D(1), Gupta S(1), Kumar S(1), Raghava GP(1).

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, 160036,

India.

Development of an effective vaccine against drug-resistant Mycobacterium

tuberculosis (Mtb) is crucial for saving millions of premature deaths every year

due to tuberculosis. This paper describes a web portal developed for assisting

researchers in designing vaccines against emerging Mtb strains using traditional

and modern approaches. Firstly, we annotated 59 genomes of Mycobacterium species

to understand similarity/dissimilarity between tuberculoid, non-tuberculoid and

vaccine strains at genome level. Secondly, antigen-based vaccine candidates have

been predicted in each Mtb strain. Thirdly, epitopes-based vaccine candidates

were predicted/discovered in above antigen-based vaccine candidates that can

stimulate all arms of immune system. Finally, a database of predicted vaccine

candidates at epitopes as well at antigen level has been developed for above

strains. In order to design vaccine against a newly sequenced genome of Mtb

strain, server integrates three modules for identification of strain-, antigen-,

epitope-specific vaccine candidates. We observed that 103,522 unique peptides

(9mers) had the potential to induce an antibody response and/or promiscuous

binder to MHC alleles and/or have the capability to stimulate T lymphocytes. In

summary, this web-portal will be useful for researchers working on designing

vaccines against Mtb including drug-resistant strains.AVAILABILITY: The database

is available freely at http://crdd.osdd.net/raghava/mtbveb/.

DOI: 10.1371/journal.pone.0153771

PMCID: PMC4838326

PMID: 27096425 [Indexed for MEDLINE]

278. BMC Bioinformatics. 2016 Apr 18;17:167. doi: 10.1186/s12859-016-1002-0.

ChemiRs: a web application for microRNAs and chemicals.

Su EC(1), Chen YS(2), Tien YC(2), Liu J(3), Ho BC(4), Yu SL(4), Singh S(5).

Author information:

(1)Graduate Institute of Biomedical Informatics, College of Medical Science and

Technology, Taipei Medical University, Taipei, 110, Taiwan. (2)Department of Life

Science, College of Science, National Taiwan Normal University, Taipei, 116,

Taiwan. (3)Department of Civil Engineering, College of Engineering, National

Taiwan University College of Engineering, Taipei, 106, Taiwan. (4)Department of

Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine,

National Taiwan University, Taipei, 100, Taiwan. (5)Department of Life Science,

College of Science, National Taiwan Normal University, Taipei, 116, Taiwan.

sher@ntnu.edu.tw.

BACKGROUND: MicroRNAs (miRNAs) are about 22 nucleotides, non-coding RNAs that

affect various cellular functions, and play a regulatory role in different

organisms including human. Until now, more than 2500 mature miRNAs in human have

been discovered and registered, but still lack of information or algorithms to

reveal the relations among miRNAs, environmental chemicals and human health.

Chemicals in environment affect our health and daily life, and some of them can

lead to diseases by inferring biological pathways.

RESULTS: We develop a creditable online web server, ChemiRs, for predicting

interactions and relations among miRNAs, chemicals and pathways. The database not

only compares gene lists affected by chemicals and miRNAs, but also incorporates

curated pathways to identify possible interactions.

CONCLUSIONS: Here, we manually retrieved associations of miRNAs and chemicals

from biomedical literature. We developed an online system, ChemiRs, which

contains miRNAs, diseases, Medical Subject Heading (MeSH) terms, chemicals,

genes, pathways and PubMed IDs. We connected each miRNA to miRBase, and every

current gene symbol to HUGO Gene Nomenclature Committee (HGNC) for genome

annotation. Human pathway information is also provided from KEGG and REACTOME

databases. Information about Gene Ontology (GO) is queried from GO Online SQL

Environment (GOOSE). With a user-friendly interface, the web application is easy

to use. Multiple query results can be easily integrated and exported as report

documents in PDF format. Association analysis of miRNAs and chemicals can help us

understand the pathogenesis of chemical components. ChemiRs is freely available

for public use at http://omics.biol.ntnu.edu.tw/ChemiRs .

DOI: 10.1186/s12859-016-1002-0

PMCID: PMC4836156

PMID: 27091357 [Indexed for MEDLINE]

279. Database (Oxford). 2016 Apr 17;2016. pii: baw056. doi: 10.1093/database/baw056.

Print 2016.

Kalium: a database of potassium channel toxins from scorpion venom.

Kuzmenkov AI(1), Krylov NA(2), Chugunov AO(1), Grishin EV(1), Vassilevski AA(3).

Author information:

(1)Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of

Sciences, Moscow 117997, Russia. (2)Shemyakin-Ovchinnikov Institute of Bioorganic

Chemistry, Russian Academy of Sciences, Moscow 117997, Russia Joint Supercomputer

Center, Russian Academy of Sciences, Moscow 119991, Russia.

(3)Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of

Sciences, Moscow 117997, Russia avas@ibch.ru.

Kalium (http://kaliumdb.org/) is a manually curated database that accumulates

data on potassium channel toxins purified from scorpion venom (KTx). This

database is an open-access resource, and provides easy access to pages of other

databases of interest, such as UniProt, PDB, NCBI Taxonomy Browser, and PubMed.

General achievements of Kalium are a strict and easy regulation of KTx

classification based on the unified nomenclature supported by researchers in the

field, removal of peptides with partial sequence and entries supported by

transcriptomic information only, classification of β-family toxins, and addition

of a novel λ-family. Molecules presented in the database can be processed by the

Clustal Omega server using a one-click option. Molecular masses of mature

peptides are calculated and available activity data are compiled for all KTx. We

believe that Kalium is not only of high interest to professional toxinologists,

but also of general utility to the scientific community.Database

URL:http://kaliumdb.org/.

© The Author(s) 2016. Published by Oxford University Press.

DOI: 10.1093/database/baw056

PMCID: PMC4834203

PMID: 27087309 [Indexed for MEDLINE]

280. Bioinformatics. 2016 Apr 15;32(8):1217-9. doi: 10.1093/bioinformatics/btv750.

Epub 2015 Dec 24.

MCAST: scanning for cis-regulatory motif clusters.

Grant CE(1), Johnson J(2), Bailey TL(2), Noble WS(3).

Author information:

(1)Department of Genome Sciences, University of Washington, Seattle, WA, USA.

(2)Institute for Molecular Bioscience, The University of Queensland, Brisbane,

Australia and. (3)Department of Genome Sciences, University of Washington,

Seattle, WA, USA, Department of Computer Science and Engineering, University of

Washington, Seattle, WA, USA.

Precise regulatory control of genes, particularly in eukaryotes, frequently

requires the joint action of multiple sequence-specific transcription factors. A

cis-regulatory module (CRM) is a genomic locus that is responsible for gene

regulation and that contains multiple transcription factor binding sites in close

proximity. Given a collection of known transcription factor binding motifs, many

bioinformatics methods have been proposed over the past 15 years for identifying

within a genomic sequence candidate CRMs consisting of clusters of those

motifs.RESULTS: The MCAST algorithm uses a hidden Markov model with a

P-value-based scoring scheme to identify candidate CRMs. Here, we introduce a new

version of MCAST that offers improved graphical output, a dynamic background

model, statistical confidence estimates based on false discovery rate estimation

and, most significantly, the ability to predict CRMs while taking into account

epigenomic data such as DNase I sensitivity or histone modification data. We

demonstrate the validity of MCAST's statistical confidence estimates and the

utility of epigenomic priors in identifying CRMs.

AVAILABILITY AND IMPLEMENTATION: MCAST is part of the MEME Suite software

toolkit. A web server and source code are available at http://meme-suite.org and

http://alternate.meme-suite.org

CONTACT: t.bailey@imb.uq.edu.au or william-noble@uw.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv750

PMCID: PMC4907379 [Available on 2017-04-15]

PMID: 26704599

281. Bioinformatics. 2016 Apr 15;32(8):1238-40. doi: 10.1093/bioinformatics/btv748.

Epub 2015 Dec 24.

Foldalign 2.5: multithreaded implementation for pairwise structural RNA

alignment.

Sundfeld D(1), Havgaard JH(2), de Melo AC(3), Gorodkin J(2).

Author information:

(1)Center for Non-Coding RNA in Technology and Health, IKVH, University of

Copenhagen, Frederiksberg, Denmark and Department of Computer Science, University

of Brasilia, Brasília, DF, Brazil. (2)Center for Non-Coding RNA in Technology and

Health, IKVH, University of Copenhagen, Frederiksberg, Denmark and. (3)Department

of Computer Science, University of Brasilia, Brasília, DF, Brazil.

MOTIVATION: Structured RNAs can be hard to search for as they often are not well

conserved in their primary structure and are local in their genomic or

transcriptomic context. Thus, the need for tools which in particular can make

local structural alignments of RNAs is only increasing.

RESULTS: To meet the demand for both large-scale screens and hands on analysis

through web servers, we present a new multithreaded version of Foldalign. We

substantially improve execution time while maintaining all previous

functionalities, including carrying out local structural alignments of sequences

with low similarity. Furthermore, the improvements allow for comparing longer

RNAs and increasing the sequence length. For example, lengths in the range

2000-6000 nucleotides improve execution up to a factor of five.

AVAILABILITY AND IMPLEMENTATION: The Foldalign software and the web server are

available at http://rth.dk/resources/foldalign

CONTACT: gorodkin@rth.dk

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv748

PMCID: PMC4824132

PMID: 26704597

282. Bioinformatics. 2016 Apr 15;32(8):1229-31. doi: 10.1093/bioinformatics/btv726.

Epub 2015 Dec 12.

INSECT 2.0: a web-server for genome-wide cis-regulatory modules prediction.

Parra RG(1), Rohr CO(1), Koile D(2), Perez-Castro C(2), Yankilevich P(2).

Author information:

(1)Facultad De Ciencias Exactas Y Naturales, Universidad De Buenos Aires, Buenos

Aires, Argentina and. (2)Instituto De Investigacion En Biomedicina De Buenos

Aires (IBioBA), CONICET, Partner Institute of the Max Planck Society, Godoy Cruz

2390, Buenos Aires C1425FQA, Argentina.

INSECT is a user-friendly web server to predict the occurrence of Cis-Regulatory

Modules (CRMs), which control gene expression. Here, we present a new release of

INSECT which includes several new features, such as whole genome analysis,

nucleosome occupancy predictions, and which provides additional links to

third-party functional tools that complement user capabilities, CRM analysis and

hypothesis construction. Improvements in the core implementation have led to a

faster and more efficient tool. In addition, this new release introduces a new

interface designed for a more integrative and dynamic user

experience.AVAILABILITY AND IMPLEMENTATION:

http://bioinformatics.ibioba-mpsp-conicet.gov.ar/INSECT2 CONTACT:

pyankilevich@ibioba-mpsp-conicet.gov.ar.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv726

PMID: 26656931

283. Bioinformatics. 2016 Apr 15;32(8):1158-62. doi: 10.1093/bioinformatics/btv709.

Epub 2015 Dec 7.

Improved topology prediction using the terminal hydrophobic helices rule.

Peters C(1), Tsirigos KD(1), Shu N(2), Elofsson A(1).

Author information:

(1)Department of Biochemistry and Biophysics, Science for Life Laboratory and.

(2)Department of Biochemistry and Biophysics, Science for Life Laboratory and

Sweden Bioinformatics Infrastructure for Life Sciences (BILS), Stockholm

University, Solna 17121, Sweden.

MOTIVATION: The translocon recognizes sufficiently hydrophobic regions of a

protein and inserts them into the membrane. Computational methods try to

determine what hydrophobic regions are recognized by the translocon. Although

these predictions are quite accurate, many methods still fail to distinguish

marginally hydrophobic transmembrane (TM) helices and equally hydrophobic regions

in soluble protein domains. In vivo, this problem is most likely avoided by

targeting of the TM-proteins, so that non-TM proteins never see the translocon.

Proteins are targeted to the translocon by an N-terminal signal peptide. The

targeting is also aided by the fact that the N-terminal helix is more hydrophobic

than other TM-helices. In addition, we also recently found that the C-terminal

helix is more hydrophobic than central helices. This information has not been

used in earlier topology predictors.

RESULTS: Here, we use the fact that the N- and C-terminal helices are more

hydrophobic to develop a new version of the first-principle-based topology

predictor, SCAMPI. The new predictor has two main advantages; first, it can be

used to efficiently separate membrane and non-membrane proteins directly without

the use of an extra prefilter, and second it shows improved performance for

predicting the topology of membrane proteins that contain large non-membrane

domains.

AVAILABILITY AND IMPLEMENTATION: The predictor, a web server and all datasets are

available at http://scampi.bioinfo.se/

CONTACT: arne@bioinfo.se

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv709

PMID: 26644416

284. PLoS One. 2016 Apr 14;11(4):e0153503. doi: 10.1371/journal.pone.0153503.

eCollection 2016.

Identification of Multi-Functional Enzyme with Multi-Label Classifier.

Che Y(1), Ju Y(1), Xuan P(2), Long R(3), Xing F(4).

Author information:

(1)School of Information Science and Technology, Xiamen University, Xiamen,

Fujian 361005, China. (2)School of Computer Science and Technology, Heilongjiang

University, Harbin 150080, China. (3)School of Computer Science and Technology,

Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong

518055, China. (4)School of Aerospace Engineering, Xiamen University, Xiamen,

Fujian 361005, China.

Enzymes are important and effective biological catalyst proteins participating in

almost all active cell processes. Identification of multi-functional enzymes is

essential in understanding the function of enzymes. Machine learning methods

perform better in protein structure and function prediction than traditional

biological wet experiments. Thus, in this study, we explore an efficient and

effective machine learning method to categorize enzymes according to their

function. Multi-functional enzymes are predicted with a special machine learning

strategy, namely, multi-label classifier. Sequence features are extracted from a

position-specific scoring matrix with autocross-covariance transformation.

Experiment results show that the proposed method obtains an accuracy rate of

94.1% in classifying six main functional classes through five cross-validation

tests and outperforms state-of-the-art methods. In addition, 91.25% accuracy is

achieved in multi-functional enzyme prediction, which is often ignored in other

enzyme function prediction studies. The online prediction server and datasets can

be accessed from the link http://server.malab.cn/MEC/.

DOI: 10.1371/journal.pone.0153503

PMCID: PMC4831692

PMID: 27078147 [Indexed for MEDLINE]

285. Nat Commun. 2016 Apr 13;7:11257. doi: 10.1038/ncomms11257.

Fast and sensitive taxonomic classification for metagenomics with Kaiju.

Menzel P(1), Ng KL(1), Krogh A(1).

Author information:

(1)Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark.

Metagenomics emerged as an important field of research not only in microbial

ecology but also for human health and disease, and metagenomic studies are

performed on increasingly larger scales. While recent taxonomic classification

programs achieve high speed by comparing genomic k-mers, they often lack

sensitivity for overcoming evolutionary divergence, so that large fractions of

the metagenomic reads remain unclassified. Here we present the novel metagenome

classifier Kaiju, which finds maximum (in-)exact matches on the protein-level

using the Burrows-Wheeler transform. We show in a genome exclusion benchmark that

Kaiju classifies reads with higher sensitivity and similar precision compared

with current k-mer-based classifiers, especially in genera that are

underrepresented in reference databases. We also demonstrate that Kaiju

classifies up to 10 times more reads in real metagenomes. Kaiju can process

millions of reads per minute and can run on a standard PC. Source code and web

server are available at http://kaiju.binf.ku.dk.

DOI: 10.1038/ncomms11257

PMCID: PMC4833860

PMID: 27071849 [Indexed for MEDLINE]

286. Int J Mol Sci. 2016 Apr 12;17(4):547. doi: 10.3390/ijms17040547.

SAAMBE: Webserver to Predict the Charge of Binding Free Energy Caused by Amino

Acids Mutations.

Petukh M(1), Dai L(2,)(3), Alexov E(4).

Author information:

(1)Computational Biophysics and Bioinformatics, Physics Department, Clemson

University, Clemson, SC 29634, USA. mpetukh@clemson.edu. (2)Computational

Biophysics and Bioinformatics, Physics Department, Clemson University, Clemson,

SC 29634, USA. luogend@g.clemson.edu. (3)Department of Computer Sciences, Clemson

University, Clemson, SC 29634, USA. luogend@g.clemson.edu. (4)Computational

Biophysics and Bioinformatics, Physics Department, Clemson University, Clemson,

SC 29634, USA. ealexov@g.clemson.edu.

Predicting the effect of amino acid substitutions on protein-protein affinity

(typically evaluated via the change of protein binding free energy) is important

for both understanding the disease-causing mechanism of missense mutations and

guiding protein engineering. In addition, researchers are also interested in

understanding which energy components are mostly affected by the mutation and how

the mutation affects the overall structure of the corresponding protein. Here we

report a webserver, the Single Amino Acid Mutation based change in Binding free

Energy (SAAMBE) webserver, which addresses the demand for tools for predicting

the change of protein binding free energy. SAAMBE is an easy to use webserver,

which only requires that a coordinate file be inputted and the user is provided

with various, but easy to navigate, options. The user specifies the mutation

position, wild type residue and type of mutation to be made. The server predicts

the binding free energy change, the changes of the corresponding energy

components and provides the energy minimized 3D structure of the wild type and

mutant proteins for download. The SAAMBE protocol performance was tested by

benchmarking the predictions against over 1300 experimentally determined changes

of binding free energy and a Pearson correlation coefficient of 0.62 was

obtained. How the predictions can be used for discriminating disease-causing from

harmless mutations is discussed. The webserver can be accessed via

http://compbio.clemson.edu/saambe\_webserver/.

DOI: 10.3390/ijms17040547

PMCID: PMC4849003

PMID: 27077847 [Indexed for MEDLINE]

287. Bioinformation. 2016 Apr 10;12(2):74-77. doi: 10.6026/97320630012074. eCollection

2016.

MFPPI - Multi FASTA ProtParam Interface.

Garg VK(1), Avashthi H(1), Tiwari A(1), Jain PA(1), Ramkete PW(2), Kayastha

AM(3), Singh VK(3).

Author information:

(1)Department of Computational Biology & Bioinformatics, Jacob School of

Biotechnology & Bioengineering, Sam Higginbottom Institute of Agriculture

Technology & Sciences, Allahabad-211007, Uttar Pradesh, Bharat (India).

(2)Department of Biological Sciences, Sam Higginbottom Institute of Agriculture

Technology & Sciences, Allahabad-211007, Uttar Pradesh, Bharat (India). (3)Centre

for Bioinformatics, School of Biotechnology, Institute of Science, Banaras Hindu

University, Varanasi-221005, Uttar Pradesh, Bharat (India).

Physico-chemical properties reflect the functional and structural characteristics

of a protein. The comparative study of the physicochemical properties is

important to know role of a protein in exploring its molecular evolution. A

number of online and offline tools are available for calculating the

physico-chemical properties of a single protein sequence. However, a tool is not

available for a comparative study with graphical visualization of Multi-FASTA

sequences. Hence, we describe the development and utility of MFPPI V.1.0 (a web

interface developed in JAVA platform) to input each FASTA sequence from

Multi-FASTA file into the ProtParam web server for the calculation of

physico-chemical properties. MFPPI V.1.0 calculates different physico-chemical

properties for a given set of proteins in a single run and saves the data in the

MSExcel sheet. Furthermore, it provides a graphical representation of protein

physico-chemical properties for analysis and visualization of data in a

user-friendly manner. Therefore, the output from the analysis helps to understand

compositional changes and functional relationship in evolution among organisms.

We have demonstrated the utility of MFPPI V.1.0 using 17 mtATP6 protein sequences

from different mammalian species. It is available for free at

http://insilicogenomics.in/mfpcalc/mfppi.html.

DOI: 10.6026/97320630012074

PMCID: PMC5237651

PMID: 28104964

288. J Theor Biol. 2016 Apr 7;394:223-30. doi: 10.1016/j.jtbi.2016.01.020. Epub 2016

Jan 22.

pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble

random forest approach.

Jia J(1), Liu Z(2), Xiao X(3), Liu B(4), Chou KC(5).

Author information:

(1)Computer Department, Jingdezhen Ceramic Institute, Jing-De-Zhen, 333403,

China; Computer science, University of Birmingham, B29 2TT, UK; Gordon Life

Science Institute, Boston, MA 02478, USA. Electronic address:

jjia@gordonlifescience.org. (2)Computer Department, Jingdezhen Ceramic Institute,

Jing-De-Zhen, 333403, China. Electronic address: liuzi189836@163.com. (3)Computer

Department, Jingdezhen Ceramic Institute, Jing-De-Zhen, 333403, China; Gordon

Life Science Institute, Boston, MA 02478, USA. Electronic address:

xxiao@gordonlifescience.org. (4)Computer Department, Jingdezhen Ceramic

Institute, Jing-De-Zhen, 333403, China. Electronic address: lbx1966@163.com.

(5)Gordon Life Science Institute, Boston, MA 02478, USA; Center of Excellence in

Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi

Arabia. Electronic address: kcchou@gordonlifescience.org.

Being one type of post-translational modifications (PTMs), protein lysine

succinylation is important in regulating varieties of biological processes. It is

also involved with some diseases, however. Consequently, from the angles of both

basic research and drug development, we are facing a challenging problem: for an

uncharacterized protein sequence having many Lys residues therein, which ones can

be succinylated, and which ones cannot? To address this problem, we have

developed a predictor called pSuc-Lys through (1) incorporating the

sequence-coupled information into the general pseudo amino acid composition, (2)

balancing out skewed training dataset by random sampling, and (3) constructing an

ensemble predictor by fusing a series of individual random forest classifiers.

Rigorous cross-validations indicated that it remarkably outperformed the existing

methods. A user-friendly web-server for pSuc-Lys has been established at

http://www.jci-bioinfo.cn/pSuc-Lys, by which users can easily obtain their

desired results without the need to go through the complicated mathematical

equations involved. It has not escaped our notice that the formulation and

approach presented here can also be used to analyze many other problems in

computational proteomics.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2016.01.020

PMID: 26807806 [Indexed for MEDLINE]

289. Nucleic Acids Res. 2016 Apr 7;44(6):e53. doi: 10.1093/nar/gkv1335. Epub 2015 Dec

3.

Prioritizing and selecting likely novel miRNAs from NGS data.

Backes C(1), Meder B(2), Hart M(3), Ludwig N(3), Leidinger P(3), Vogel B(2),

Galata V(1), Roth P(4), Menegatti J(5), Grässer F(5), Ruprecht K(6), Kahraman

M(1), Grossmann T(1), Haas J(2), Meese E(3), Keller A(7).

Author information:

(1)Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany.

(2)Internal Medicine II, University Hospital Heidelberg, Heidelberg, Germany.

(3)Department of Human Genetics, Saarland University, Homburg, Germany.

(4)University Hospital Zurich, Department of Neurology and University of Zurich,

Switzerland. (5)Department of Virology, Saarland University Medical School,

Homburg, Germany. (6)Department of Neurology, Charité Universitätsmedizin Berlin,

Germany. (7)Chair for Clinical Bioinformatics, Saarland University, Saarbrücken,

Germany andreas.keller@ccb.uni-saarland.de.

Small non-coding RNAs play a key role in many physiological and pathological

processes. Since 2004, miRNA sequences have been catalogued in miRBase, which is

currently in its 21st version. We investigated sequence and structural features

of miRNAs annotated in the miRBase and compared them between different versions

of this reference database. We have identified that the two most recent releases

(v20 and v21) are influenced by next-generation sequencing based miRNA

predictions and show significant deviation from miRNAs discovered prior to the

high-throughput profiling period. From the analysis of miRBase, we derived a set

of key characteristics to predict new miRNAs and applied the implemented

algorithm to evaluate novel blood-borne miRNA candidates. We carried out 705

individual whole miRNA sequencings of blood cells and collected a total of 9.7

billion reads. Using miRDeep2 we initially predicted 1452 potentially novel

miRNAs. After excluding false positives, 518 candidates remained. These novel

candidates were ranked according to their distance to the features in the early

miRBase versions allowing for an easier selection of a subset of putative miRNAs

for validation. Selected candidates were successfully validated by qRT-PCR and

northern blotting. In addition, we implemented a web-server for ranking potential

miRNA candidates, which is available at:www.ccb.uni-saarland.de/novomirank.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv1335

PMCID: PMC4824081

PMID: 26635395 [Indexed for MEDLINE]

290. BMC Bioinformatics. 2016 Apr 4;17:152. doi: 10.1186/s12859-016-0996-7.

Mutanalyst, an online tool for assessing the mutational spectrum of epPCR

libraries with poor sampling.

Ferla MP(1,)(2).

Author information:

(1)Formerly Department of Biochemistry, University of Otago, Dunedin, New

Zealand. matteo.ferla@gmail.com. (2)Present address: Biosyntia, DTU Centre for

Biosustainability, Hørsholm, Denmark. matteo.ferla@gmail.com.

BACKGROUND: Assessing library diversity is an important control step in a

directed evolution experiment. To do this, a limited amount of colonies from a

test library are sequenced and tested. In the case of an error-prone PCR library,

the spectrum of the identified mutations - the proportions of mutations of a

specific nucleobase to another- is calculated enabling the user to make more

informed predictions on library diversity and coverage. However, the calculations

of the mutational spectrum are severely affected by the limited sample sizes.

RESULTS: Here an online program, called Mutanalyst, is presented, which not only

automates the calculations, but also estimates errors involved. Specifically, the

errors are calculated thanks to the complementarity of DNA, which means that a

mutation has a complementary mutation on the other sequence. Additionally, in the

case of determining the mean number of mutations per sequence it does so by

fitting to a Poisson distribution, which is more robust than calculating the

average in light of the small sampling size.

CONCLUSION: As a result of the added measures to keep into account of small

sample size the user can better assess whether the library is satisfactory or

whether error-prone PCR conditions should be adjusted. The program is available

at www.mutanalyst.com .

DOI: 10.1186/s12859-016-0996-7

PMCID: PMC4820924

PMID: 27044645 [Indexed for MEDLINE]

291. Sci Rep. 2016 Apr 4;6:23990. doi: 10.1038/srep23990.

Protein single-model quality assessment by feature-based probability density

functions.

Cao R(1), Cheng J(1,)(2,)(3).

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, MO 65211,

USA. (2)Informatics Institute, University of Missouri, Columbia, MO 65211, USA.

(3)Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA.

Protein quality assessment (QA) has played an important role in protein structure

prediction. We developed a novel single-model quality assessment method-Qprob.

Qprob calculates the absolute error for each protein feature value against the

true quality scores (i.e. GDT-TS scores) of protein structural models, and uses

them to estimate its probability density distribution for quality assessment.

Qprob has been blindly tested on the 11th Critical Assessment of Techniques for

Protein Structure Prediction (CASP11) as MULTICOM-NOVEL server. The official CASP

result shows that Qprob ranks as one of the top single-model QA methods. In

addition, Qprob makes contributions to our protein tertiary structure predictor

MULTICOM, which is officially ranked 3rd out of 143 predictors. The good

performance shows that Qprob is good at assessing the quality of models of hard

targets. These results demonstrate that this new probability density distribution

based method is effective for protein single-model quality assessment and is

useful for protein structure prediction. The webserver of Qprob is available at:

http://calla.rnet.missouri.edu/qprob/. The software is now freely available in

the web server of Qprob.

DOI: 10.1038/srep23990

PMCID: PMC4819172

PMID: 27041353 [Indexed for MEDLINE]

292. Bioinformatics. 2016 Apr 1;32(7):1091-3. doi: 10.1093/bioinformatics/btv705. Epub

2015 Dec 1.

LedPred: an R/bioconductor package to predict regulatory sequences using support

vector machines.

Seyres D(1), Darbo E(2), Perrin L(3), Herrmann C(4), González A(1).

Author information:

(1)INSERM, UMR1090 TAGC, Marseille, F-13288 France, Aix-Marseille Université,

UMR1090 TAGC, Marseille, F-13288 France. (2)Cancer Research UK, London Research

Institute, London WC2A 3LY, UK. (3)INSERM, UMR1090 TAGC, Marseille, F-13288

France, Aix-Marseille Université, UMR1090 TAGC, Marseille, F-13288 France, CNRS,

Marseille, France and. (4)IPMB, Universität Heidelberg and Department of

Theoretical Bioinformatics, DKFZ, Heidelberg 69120, Germany.

Supervised classification based on support vector machines (SVMs) has

successfully been used for the prediction of cis-regulatory modules (CRMs).

However, no integrated tool using such heterogeneous data as position-specific

scoring matrices, ChIP-seq data or conservation scores is currently available.

Here, we present LedPred, a flexible SVM workflow that predicts new regulatory

sequences based on the annotation of known CRMs, which are associated to a large

variety of feature types. LedPred is provided as an R/Bioconductor package

connected to an online server to avoid installation of non-R software. Due to the

heterogeneous CRM feature integration, LedPred excels at the prediction of

regulatory sequences in Drosophila and mouse datasets compared with similar

SVM-based software.AVAILABILITY AND IMPLEMENTATION: LedPred is available on

GitHub: https://github.com/aitgon/LedPred and Bioconductor:

http://bioconductor.org/packages/release/bioc/html/LedPred.html under the MIT

license.

CONTACT: aitor.gonzalez@univ-amu.fr

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv705

PMID: 26628586

293. Bioinformatics. 2016 Apr 1;32(7):1083-4. doi: 10.1093/bioinformatics/btv689. Epub

2015 Nov 24.

Synchronized navigation and comparative analyses across Ensembl complete

bacterial genomes with INSYGHT.

Lacroix T(1), Thérond S(2), Rugeri M(2), Nicolas P(1), Gendrault A(1), Loux V(1),

Gibrat JF(1).

Author information:

(1)INRA, UR1404, Unité Mathématiques et Informatique Appliquées du Génome à

l'Environnement, 78350 Jouy-en-Josas, France and. (2)IDRIS-CNRS, UPS851, 91400

Orsay, France.

MOTIVATION: High-throughput sequencing technologies provide access to an

increasing number of bacterial genomes. Today, many analyses involve the

comparison of biological properties among many strains of a given species, or

among species of a particular genus. Tools that can help the microbiologist with

these tasks become increasingly important.

RESULTS: Insyght is a comparative visualization tool whose core features combine

a synchronized navigation across genomic data of multiple organisms with a

versatile interoperability between complementary views. In this work, we have

greatly increased the scope of the Insyght public dataset by including 2688

complete bacterial genomes available in Ensembl thus vastly improving its

phylogenetic coverage. We also report the development of a virtual machine that

allows users to easily set up and customize their own local Insyght server.

AVAILABILITY AND IMPLEMENTATION: http://genome.jouy.inra.fr/Insyght

CONTACT: Thomas.Lacroix@jouy.inra.fr.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv689

PMCID: PMC4896367

PMID: 26607491

294. Comput Biol Med. 2016 Apr 1;71:156-61. doi: 10.1016/j.compbiomed.2016.02.012.

Epub 2016 Feb 26.

Predicting bacteriophage proteins located in host cell with feature selection

technique.

Ding H(1), Liang ZY(2), Guo FB(2), Huang J(2), Chen W(3), Lin H(4).

Author information:

(1)Key Laboratory for NeuroInformation of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology and Center for Information

in Biomedicine, University of Electronic Science and Technology of China, Chengdu

610054, China. Electronic address: hding@uestc.edu.cn. (2)Key Laboratory for

NeuroInformation of Ministry of Education, Center of Bioinformatics, School of

Life Science and Technology and Center for Information in Biomedicine, University

of Electronic Science and Technology of China, Chengdu 610054, China. (3)Key

Laboratory for NeuroInformation of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology and Center for Information

in Biomedicine, University of Electronic Science and Technology of China, Chengdu

610054, China; Department of Physics, School of Sciences, and Center for Genomics

and Computational Biology, North China University of Science and Technology,

Tangshan 063000, China. Electronic address: greatchen@heuu.edu.cn. (4)Key

Laboratory for NeuroInformation of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology and Center for Information

in Biomedicine, University of Electronic Science and Technology of China, Chengdu

610054, China. Electronic address: hlin@uestc.edu.cn.

A bacteriophage is a virus that can infect a bacterium. The fate of an infected

bacterium is determined by the bacteriophage proteins located in the host cell.

Thus, reliably identifying bacteriophage proteins located in the host cell is

extremely important to understand their functions and discover potential

anti-bacterial drugs. Thus, in this paper, a computational method was developed

to recognize bacteriophage proteins located in host cells based only on their

amino acid sequences. The analysis of variance (ANOVA) combined with incremental

feature selection (IFS) was proposed to optimize the feature set. Using a

jackknife cross-validation, our method can discriminate between bacteriophage

proteins located in a host cell and the bacteriophage proteins not located in a

host cell with a maximum overall accuracy of 84.2%, and can further classify

bacteriophage proteins located in host cell cytoplasm and in host cell membranes

with a maximum overall accuracy of 92.4%. To enhance the value of the practical

applications of the method, we built a web server called PHPred

(〈http://lin.uestc.edu.cn/server/PHPred〉). We believe that the PHPred will become

a powerful tool to study bacteriophage proteins located in host cells and to

guide related drug discovery.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiomed.2016.02.012

PMID: 26945463 [Indexed for MEDLINE]

295. Int J Cancer. 2016 Apr 1;138(7):1765-76. doi: 10.1002/ijc.29897.

DrugTargetInspector: An assistance tool for patient treatment stratification.

Schneider L(1,)(2), Stöckel D(1,)(2), Kehl T(1,)(2), Gerasch A(3), Ludwig N(4),

Leidinger P(4), Huwer H(5), Tenzer S(6), Kohlbacher O(3,)(7), Hildebrandt A(8),

Kaufmann M(3), Gessler M(9), Keller A(1), Meese E(4), Graf N(10), Lenhof HP(1).

Author information:

(1)Center for Bioinformatics, Saarland University, Saarbrücken, Germany.

(2)Saarbrücken Graduate School of Computer Science, Saarland University,

Saarbrücken, Germany. (3)Center for Bioinformatics, Eberhard-Karls-University,

Tübingen, Germany. (4)Department of Human Genetics, Medical School, Saarland

University, Homburg, Germany. (5)Department of Cardiothoracic Surgery, Völklingen

Heart Center, Völklingen, Germany. (6)Institute for Immunology, Johannes

Gutenberg University Mainz, Mainz, Germany. (7)Biomolecular Interactions, Max

Planck Institute for Developmental Biology, Tübingen, Germany. (8)Institute of

Computer Science, Johannes Gutenberg University Mainz, Mainz, Germany.

(9)Theodor-Boveri-Institute/Biocenter, Developmental Biochemistry, and

Comprehensive Cancer Center Mainfranken, Würzburg University, Würzburg, Germany.

(10)Department of Pediatric Oncology and Hematology, Medical School, Saarland

University, Homburg, Germany.

Cancer is a large class of diseases that are characterized by a common set of

features, known as the Hallmarks of cancer. One of these hallmarks is the

acquisition of genome instability and mutations. This, combined with high

proliferation rates and failure of repair mechanisms, leads to clonal evolution

as well as a high genotypic and phenotypic diversity within the tumor. As a

consequence, treatment and therapy of malignant tumors is still a grand

challenge. Moreover, under selective pressure, e.g., caused by chemotherapy,

resistant subpopulations can emerge that then may lead to relapse. In order to

minimize the risk of developing multidrug-resistant tumor cell populations,

optimal (combination) therapies have to be determined on the basis of an in-depth

characterization of the tumor's genetic and phenotypic makeup, a process that is

an important aspect of stratified medicine and precision medicine. We present

DrugTargetInspector (DTI), an interactive assistance tool for treatment

stratification. DTI analyzes genomic, transcriptomic, and proteomic datasets and

provides information on deregulated drug targets, enriched biological pathways,

and deregulated subnetworks, as well as mutations and their potential effects on

putative drug targets and genes of interest. To demonstrate DTI's broad scope of

applicability, we present case studies on several cancer types and different

types of input -omics data. DTI's integrative approach allows users to

characterize the tumor under investigation based on various -omics datasets and

to elucidate putative treatment options based on clinical decision guidelines,

but also proposing additional points of intervention that might be neglected

otherwise. DTI can be freely accessed at http://dti.bioinf.uni-sb.de.

© 2015 UICC.

DOI: 10.1002/ijc.29897

PMID: 26501925 [Indexed for MEDLINE]

296. Mol Biosyst. 2016 Apr;12(4):1269-75. doi: 10.1039/c5mb00883b. Epub 2016 Feb 17.

Identification of immunoglobulins using Chou's pseudo amino acid composition with

feature selection technique.

Tang H(1), Chen W, Lin H.

Author information:

(1)Department of Pathophysiology, Sichuan Medical University, Luzhou 646000,

China. tanghua771211@aliyun.com.

Immunoglobulins, also called antibodies, are a group of cell surface proteins

which are produced by the immune system in response to the presence of a foreign

substance (called antigen). They play key roles in many medical, diagnostic and

biotechnological applications. Correct identification of immunoglobulins is

crucial to the comprehension of humoral immune function. With the avalanche of

protein sequences identified in postgenomic age, it is highly desirable to

develop computational methods to timely identify immunoglobulins. In view of

this, we designed a predictor called "IGPred" by formulating protein sequences

with the pseudo amino acid composition into which nine physiochemical properties

of amino acids were incorporated. Jackknife cross-validated results showed that

96.3% of immunoglobulins and 97.5% of non-immunoglobulins can be correctly

predicted, indicating that IGPred holds very high potential to become a useful

tool for antibody analysis. For the convenience of most experimental scientists,

a web-server for IGPred was established at http://lin.uestc.edu.cn/server/IGPred.

We believe that the web-server will become a powerful tool to study

immunoglobulins and to guide related experimental validations.

DOI: 10.1039/c5mb00883b

PMID: 26883492 [Indexed for MEDLINE]

297. Oncotarget. 2016 Mar 29;7(13):16895-909. doi: 10.18632/oncotarget.7815.

iACP: a sequence-based tool for identifying anticancer peptides.

Chen W(1,)(2), Ding H(3), Feng P(4), Lin H(3,)(2), Chou KC(2,)(5).

Author information:

(1)Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, North China University of Science and Technology,

Tangshan, China. (2)Gordon Life Science Institute, Belmont, Massachusetts, United

States of America. (3)Key Laboratory for Neuro-Information of Ministry of

Education, Center of Bioinformatics, School of Life Science and Technology,

University of Electronic Science and Technology of China, Chengdu, China.

(4)School of Public Health, North China University of Science and Technology,

Tangshan, China. (5)Center of Excellence in Genomic Medicine Research (CEGMR),

King Abdulaziz University, Jeddah, Saudi Arabia.

Cancer remains a major killer worldwide. Traditional methods of cancer treatment

are expensive and have some deleterious side effects on normal cells.

Fortunately, the discovery of anticancer peptides (ACPs) has paved a new way for

cancer treatment. With the explosive growth of peptide sequences generated in the

post genomic age, it is highly desired to develop computational methods for

rapidly and effectively identifying ACPs, so as to speed up their application in

treating cancer. Here we report a sequence-based predictor called iACP developed

by the approach of optimizing the g-gap dipeptide components. It was demonstrated

by rigorous cross-validations that the new predictor remarkably outperformed the

existing predictors for the same purpose in both overall accuracy and stability.

For the convenience of most experimental scientists, a publicly accessible

web-server for iACP has been established at http://lin.uestc.edu.cn/server/iACP,

by which users can easily obtain their desired results.

DOI: 10.18632/oncotarget.7815

PMCID: PMC4941358

PMID: 26942877

298. Database (Oxford). 2016 Mar 28;2016. pii: baw037. doi: 10.1093/database/baw037.

Print 2016.

myPhyloDB: a local web server for the storage and analysis of metagenomic data.

Manter DK(1), Korsa M(2), Tebbe C(2), Delgado JA(2).

Author information:

(1)USDA-ARS, Soil-Plant-Nutrient-Research, Fort Collins, CO 80526, USA

Daniel.manter@ars.usda.gov. (2)USDA-ARS, Soil-Plant-Nutrient-Research, Fort

Collins, CO 80526, USA.

myPhyloDB v.1.1.2 is a user-friendly personal database with a browser-interface

designed to facilitate the storage, processing, analysis, and distribution of

microbial community populations (e.g. 16S metagenomics data). MyPhyloDB archives

raw sequencing files, and allows for easy selection of project(s)/sample(s) of

any combination from all available data in the database. The data processing

capabilities of myPhyloDB are also flexible enough to allow the upload and

storage of pre-processed data, or use the built-in Mothur pipeline to automate

the processing of raw sequencing data. myPhyloDB provides several analytical

(e.g. analysis of covariance,t-tests, linear regression, differential abundance

(DESeq2), and principal coordinates analysis (PCoA)) and normalization

(rarefaction, DESeq2, and proportion) tools for the comparative analysis of

taxonomic abundance, species richness and species diversity for projects of

various types (e.g. human-associated, human gut microbiome, air, soil, and water)

for any taxonomic level(s) desired. Finally, since myPhyloDB is a local

web-server, users can quickly distribute data between colleagues and end-users by

simply granting others access to their personal myPhyloDB database. myPhyloDB is

available athttp://www.ars.usda.gov/services/software/download.htm?softwareid=472

and more information along with tutorials can be found on our

websitehttp://www.myphylodb.org. Database URL:http://www.myphylodb.org.

Published by Oxford University Press 2016. This work is written by US Government

employees and is in the public domain in the United States.

DOI: 10.1093/database/baw037

PMCID: PMC4809264

PMID: 27022159 [Indexed for MEDLINE]

299. Genet Mol Res. 2016 Mar 28;15(1). doi: 10.4238/gmr.15016861.

miRQuest: integration of tools on a Web server for microRNA research.

Aguiar RR(1), Ambrosio LA(1), Sepúlveda-Hermosilla G(2), Maracaja-Coutinho

V(2,)(3,)(4), Paschoal AR(1).

Author information:

(1)Universidade Tecnológica Federal do Paraná, Cornélio Procópio, PR, Brasil.

(2)Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor,

Santiago, Chile. (3)Beagle Bioniformatics, Santiago, Chile. (4)Instituto

Vandique, João Pessoa, PB, Brasil.

This report describes the miRQuest - a novel middleware available in a Web server

that allows the end user to do the miRNA research in a user-friendly way. It is

known that there are many prediction tools for microRNA (miRNA) identification

that use different programming languages and methods to realize this task. It is

difficult to understand each tool and apply it to diverse datasets and organisms

available for miRNA analysis. miRQuest can easily be used by biologists and

researchers with limited experience with bioinformatics. We built it using the

middleware architecture on a Web platform for miRNA research that performs two

main functions: i) integration of different miRNA prediction tools for miRNA

identification in a user-friendly environment; and ii) comparison of these

prediction tools. In both cases, the user provides sequences (in FASTA format) as

an input set for the analysis and comparisons. All the tools were selected on the

basis of a survey of the literature on the available tools for miRNA prediction.

As results, three different cases of use of the tools are also described, where

one is the miRNA identification analysis in 30 different species. Finally,

miRQuest seems to be a novel and useful tool; and it is freely available for both

benchmarking and miRNA identification at

http://mirquest.integrativebioinformatics.me/.

DOI: 10.4238/gmr.15016861

PMID: 27050998 [Indexed for MEDLINE]

300. J Mol Biol. 2016 Mar 27;428(6):1394-405. doi: 10.1016/j.jmb.2016.01.012. Epub

2016 Jan 22.

EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with

Feature-Based Multiple Models.

Folkman L(1), Stantic B(2), Sattar A(3), Zhou Y(4).

Author information:

(1)Institute for Integrated and Intelligent Systems, Griffith University, 170

Kessels Road, Brisbane, Queensland 4111, Australia; Queensland Research

Laboratory, NICTA, National ICT Australia, 70-72 Bowen Street, Spring Hill,

Queensland 4000, Australia. Electronic address: lukas.folkman@griffithuni.edu.au.

(2)Institute for Integrated and Intelligent Systems, Griffith University, 170

Kessels Road, Brisbane, Queensland 4111, Australia. Electronic address:

b.stantic@griffith.edu.au. (3)Institute for Integrated and Intelligent Systems,

Griffith University, 170 Kessels Road, Brisbane, Queensland 4111, Australia;

Queensland Research Laboratory, NICTA, National ICT Australia, 70-72 Bowen

Street, Spring Hill, Queensland 4000, Australia. Electronic address:

a.sattar@griffith.edu.au. (4)Institute for Integrated and Intelligent Systems,

Griffith University, 170 Kessels Road, Brisbane, Queensland 4111, Australia;

Institute for Glycomics, Griffith University, Parklands Drive, Southport,

Queensland 4222, Australia. Electronic address: yaoqi.zhou@griffith.edu.au.

Protein engineering and characterisation of non-synonymous single nucleotide

variants (SNVs) require accurate prediction of protein stability changes (ΔΔGu)

induced by single amino acid substitutions. Here, we have developed a new

prediction method called Evolutionary, Amino acid, and Structural Encodings with

Multiple Models (EASE-MM), which comprises five specialised support vector

machine (SVM) models and makes the final prediction from a consensus of two

models selected based on the predicted secondary structure and accessible surface

area of the mutated residue. The new method is applicable to single-domain

monomeric proteins and can predict ΔΔGu with a protein sequence and mutation as

the only inputs. EASE-MM yielded a Pearson correlation coefficient of 0.53-0.59

in 10-fold cross-validation and independent testing and was able to outperform

other sequence-based methods. When compared to structure-based energy functions,

EASE-MM achieved a comparable or better performance. The application to a large

dataset of human germline non-synonymous SNVs showed that the disease-causing

variants tend to be associated with larger magnitudes of ΔΔGu predicted with

EASE-MM. The EASE-MM web-server is available at

http://sparks-lab.org/server/ease.

Copyright © 2016 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jmb.2016.01.012

PMID: 26804571 [Indexed for MEDLINE]

301. Biol Direct. 2016 Mar 25;11(1):14. doi: 10.1186/s13062-016-0118-5.

A web server for analysis, comparison and prediction of protein ligand binding

sites.

Singh H(1), Srivastava HK(1), Raghava GP(2,)(3).

Author information:

(1)Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh,

160036, India. (2)Bioinformatics Centre, CSIR-Institute of Microbial Technology,

Chandigarh, 160036, India. raghava@imtech.res.in. (3), . raghava@imtech.res.in.

BACKGROUND: One of the major challenges in the field of system biology is to

understand the interaction between a wide range of proteins and ligands. In the

past, methods have been developed for predicting binding sites in a protein for a

limited number of ligands.

RESULTS: In order to address this problem, we developed a web server named

'LPIcom' to facilitate users in understanding protein-ligand interaction.

Analysis, comparison and prediction modules are available in the "LPIcom' server

to predict protein-ligand interacting residues for 824 ligands. Each ligand must

have at least 30 protein binding sites in PDB. Analysis module of the server can

identify residues preferred in interaction and binding motif for a given ligand;

for example residues glycine, lysine and arginine are preferred in ATP binding

sites. Comparison module of the server allows comparing protein-binding sites of

multiple ligands to understand the similarity between ligands based on their

binding site. This module indicates that ATP, ADP and GTP ligands are in the same

cluster and thus their binding sites or interacting residues exhibit a high level

of similarity. Propensity-based prediction module has been developed for

predicting ligand-interacting residues in a protein for more than 800 ligands. In

addition, a number of web-based tools have been integrated to facilitate users in

creating web logo and two-sample between ligand interacting and non-interacting

residues.

CONCLUSIONS: In summary, this manuscript presents a web-server for analysis of

ligand interacting residue. This server is available for public use from URL

http://crdd.osdd.net/raghava/lpicom .

DOI: 10.1186/s13062-016-0118-5

PMCID: PMC4807588

PMID: 27016210 [Indexed for MEDLINE]

302. BMJ Open. 2016 Mar 24;6(3):e010579. doi: 10.1136/bmjopen-2015-010579.

EHDViz: clinical dashboard development using open-source technologies.

Badgeley MA(1), Shameer K(1), Glicksberg BS(1), Tomlinson MS(1), Levin MA(2),

McCormick PJ(3), Kasarskis A(4), Reich DL(3), Dudley JT(5).

Author information:

(1)Harris Center for Precision Wellness, Icahn School of Medicine at Mount Sinai,

Mount Sinai Health System, New York City, New York, USA Department of Genetics

and Genomic Sciences, Icahn Institute for Genomics and Multiscale Biology, Icahn

School of Medicine at Mount Sinai, Mount Sinai Health System, New York City, New

York, USA. (2)Department of Genetics and Genomic Sciences, Icahn Institute for

Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, Mount

Sinai Health System, New York City, New York, USA Department of Anesthesiology,

Icahn School of Medicine at Mount Sinai, Mount Sinai Health System, New York

City, New York, USA. (3)Department of Anesthesiology, Icahn School of Medicine at

Mount Sinai, Mount Sinai Health System, New York City, New York, USA.

(4)Department of Genetics and Genomic Sciences, Icahn Institute for Genomics and

Multiscale Biology, Icahn School of Medicine at Mount Sinai, Mount Sinai Health

System, New York City, New York, USA. (5)Harris Center for Precision Wellness,

Icahn School of Medicine at Mount Sinai, Mount Sinai Health System, New York

City, New York, USA Department of Genetics and Genomic Sciences, Icahn Institute

for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai,

Mount Sinai Health System, New York City, New York, USA Department of Population

Health Science and Policy, Icahn School of Medicine at Mount Sinai, Mount Sinai

Health System, New York City, New York, USA.

OBJECTIVE: To design, develop and prototype clinical dashboards to integrate

high-frequency health and wellness data streams using interactive and real-time

data visualisation and analytics modalities.

MATERIALS AND METHODS: We developed a clinical dashboard development framework

called electronic healthcare data visualization (EHDViz) toolkit for generating

web-based, real-time clinical dashboards for visualising heterogeneous

biomedical, healthcare and wellness data. The EHDViz is an extensible toolkit

that uses R packages for data management, normalisation and producing

high-quality visualisations over the web using R/Shiny web server architecture.

We have developed use cases to illustrate utility of EHDViz in different

scenarios of clinical and wellness setting as a visualisation aid for improving

healthcare delivery.

RESULTS: Using EHDViz, we prototyped clinical dashboards to demonstrate the

contextual versatility of EHDViz toolkit. An outpatient cohort was used to

visualise population health management tasks (n=14,221), and an inpatient cohort

was used to visualise real-time acuity risk in a clinical unit (n=445), and a

quantified-self example using wellness data from a fitness activity monitor worn

by a single individual was also discussed (n-of-1). The back-end system retrieves

relevant data from data source, populates the main panel of the application and

integrates user-defined data features in real-time and renders output using

modern web browsers. The visualisation elements can be customised using health

features, disease names, procedure names or medical codes to populate the

visualisations. The source code of EHDViz and various prototypes developed using

EHDViz are available in the public domain at http://ehdviz.dudleylab.org.

CONCLUSIONS: Collaborative data visualisations, wellness trend predictions, risk

estimation, proactive acuity status monitoring and knowledge of complex disease

indicators are essential components of implementing data-driven precision

medicine. As an open-source visualisation framework capable of integrating health

assessment, EHDViz aims to be a valuable toolkit for rapid design, development

and implementation of scalable clinical data visualisation dashboards.

Published by the BMJ Publishing Group Limited. For permission to use (where not

already granted under a licence) please go to

http://www.bmj.com/company/products-services/rights-and-licensing/

DOI: 10.1136/bmjopen-2015-010579

PMCID: PMC4809078

PMID: 27013597 [Indexed for MEDLINE]

303. Sci Rep. 2016 Mar 24;6:23700. doi: 10.1038/srep23700.

miR-isomiRExp: a web-server for the analysis of expression of miRNA at the

miRNA/isomiR levels.

Guo L(1), Yu J(2), Liang T(3), Zou Q(4).

Author information:

(1)Department of Bioinformatics, School of Geographic and Biologic Information,

Nanjing University of Posts and Telecommunications, Nanjing, 210023, China.

(2)Shandong Provincial Key Laboratory of Functional Macromolecular Biophysics,

Institute of Biophysics, Dezhou University, Dezhou, 253023, China. (3)Jiangsu Key

Laboratory for Molecular and Medical Biotechnology, College of Life Science,

Nanjing Normal University, Nanjing, 210023, China. (4)School of Computer Science

and Technology, Tianjin University, Tianjin, 300072, China.

MicroRNA (miRNA) locus has been found that can generate a series of varied isomiR

sequences. Most studies always focus on determining miRNA level, however, the

canonical miRNA sequence is only a specific member in the multiple isomiRs. Some

studies have shown that isomiR sequences play versatile roles in biological

progress, and the analysis and research should be simultaneously performed at the

miRNA/isomiR levels. Based on the biological characteristics of miRNA and isomiR,

we developed miR-isomiRExp to analyze expression pattern of miRNA at the

miRNA/isomiR levels, provide insights into tracking miRNA/isomiR maturation and

processing mechanisms, and reveal functional characteristics of miRNA/isomiR.

Simultaneously, we also performed expression analysis of specific human diseases

using public small RNA sequencing datasets based on the analysis platform, which

may help in surveying the potential deregulated miRNA/isomiR expression profiles,

especially sequence and function-related isomiRs for further interaction analysis

and study. The miR-isomiRExp platform provides miRNA/isomiR expression patterns

and more information to study deregulated miRNA loci and detailed isomiR

sequences. This comprehensive analysis will enrich experimental miRNA studies.

miR-isomiRExp is available at http://mirisomirexp.aliapp.com.

DOI: 10.1038/srep23700

PMCID: PMC4806314

PMID: 27009551 [Indexed for MEDLINE]

304. Sci Rep. 2016 Mar 22;6:23510. doi: 10.1038/srep23510.

DephosSite: a machine learning approach for discovering phosphotase-specific

dephosphorylation sites.

Wang X(1), Yan R(2), Song J(3,)(4,)(5).

Author information:

(1)School of Mathematics and Computer Science, Shanxi Normal University, Linfen

041004, China. (2)Institute of Applied Genomics, School of Biological Sciences

and Engineering, Fuzhou University, Fuzhou 350002, China. (3)Infection and

Immunity Program and The Department of Biochemistry and Molecular Biology,

Biomedicine Discovery Institute, Monash University, Clayton, VIC 3800, Australia.

(4)Monash Centre for Data Science, Faculty of Information Technology, Monash

University, Clayton, VIC 3800, Australia. (5)National Engineering Laboratory for

Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin

Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin,

300308, China.

Protein dephosphorylation, which is an inverse process of phosphorylation, plays

a crucial role in a myriad of cellular processes, including mitotic cycle,

proliferation, differentiation, and cell growth. Compared with tyrosine kinase

substrate and phosphorylation site prediction, there is a paucity of studies

focusing on computational methods of predicting protein tyrosine phosphatase

substrates and dephosphorylation sites. In this work, we developed two elegant

models for predicting the substrate dephosphorylation sites of three specific

phosphatases, namely, PTP1B, SHP-1, and SHP-2. The first predictor is called

MGPS-DEPHOS, which is modified from the GPS (Group-based Prediction System)

algorithm with an interpretable capability. The second predictor is called

CKSAAP-DEPHOS, which is built through the combination of support vector machine

(SVM) and the composition of k-spaced amino acid pairs (CKSAAP) encoding scheme.

Benchmarking experiments using jackknife cross validation and 30 repeats of

5-fold cross validation tests show that MGPS-DEPHOS and CKSAAP-DEPHOS achieved

AUC values of 0.921, 0.914 and 0.912, for predicting dephosphorylation sites of

the three phosphatases PTP1B, SHP-1, and SHP-2, respectively. Both methods

outperformed the previously developed kNN-DEPHOS algorithm. In addition, a web

server implementing our algorithms is publicly available at

http://genomics.fzu.edu.cn/dephossite/ for the research community.

DOI: 10.1038/srep23510

PMCID: PMC4802303

PMID: 27002216 [Indexed for MEDLINE]

305. Cell Death Dis. 2016 Mar 17;7:e2148. doi: 10.1038/cddis.2016.42.

p53MutaGene: an online tool to estimate the effect of p53 mutational status on

gene regulation in cancer.

Amelio I(1), Knight RA(1), Lisitsa A(2), Melino G(1,)(3), Antonov AV(1).

Author information:

(1)Medical Research Council Toxicology Unit, Leicester LE1 9HN, UK. (2)Institute

of Biomedical Chemistry of the Russian Academy of Medical Sciences, Moscow,

Russia. (3)Department of Experimental Medicine & Surgery, University of Rome Tor

Vergata, Rome 00133, Italy.

p53MutaGene is the first online tool for statistical validation of hypotheses

regarding the effect of p53 mutational status on gene regulation in cancer. This

tool is based on several large-scale clinical gene expression data sets and

currently covers breast, colon and lung cancers. The tool detects differential

co-expression patterns in expression data between p53 mutated versus p53 normal

samples for the user-specified genes. Statistically significant differential

co-expression for a gene pair is indicative that regulation of two genes is

sensitive to the presence of p53 mutations. p53MutaGene can be used in 'single

mode' where the user can test a specific pair of genes or in 'discovery mode'

designed for analysis of several genes. Using several examples, we demonstrate

that p53MutaGene is a useful tool for fast statistical validation in clinical

data of p53-dependent gene regulation patterns. The tool is freely available at

http://www.bioprofiling.de/tp53.

DOI: 10.1038/cddis.2016.42

PMCID: PMC4823943

PMID: 26986515 [Indexed for MEDLINE]

306. Database (Oxford). 2016 Mar 17;2016. pii: baw024. doi: 10.1093/database/baw024.

Print 2016.

dbWGFP: a database and web server of human whole-genome single nucleotide

variants and their functional predictions.

Wu J(1), Wu M(1), Li L(1), Liu Z(1), Zeng W(1), Jiang R(2).

Author information:

(1)MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for

Synthetic & Systems Biology, TNLIST, Department of Automation, Tsinghua

University, Beijing 100084, China. (2)MOE Key Laboratory of Bioinformatics,

Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST,

Department of Automation, Tsinghua University, Beijing 100084, China

ruijiang@tsinghua.edu.cn.

The recent advancement of the next generation sequencing technology has enabled

the fast and low-cost detection of all genetic variants spreading across the

entire human genome, making the application of whole-genome sequencing a tendency

in the study of disease-causing genetic variants. Nevertheless, there still lacks

a repository that collects predictions of functionally damaging effects of human

genetic variants, though it has been well recognized that such predictions play a

central role in the analysis of whole-genome sequencing data. To fill this gap,

we developed a database named dbWGFP (a database and web server of human

whole-genome single nucleotide variants and their functional predictions) that

contains functional predictions and annotations of nearly 8.58 billion possible

human whole-genome single nucleotide variants. Specifically, this database

integrates 48 functional predictions calculated by 17 popular computational

methods and 44 valuable annotations obtained from various data sources.

Standalone software, user-friendly query services and free downloads of this

database are available at http://bioinfo.au.tsinghua.edu.cn/dbwgfp. dbWGFP

provides a valuable resource for the analysis of whole-genome sequencing, exome

sequencing and SNP array data, thereby complementing existing data sources and

computational resources in deciphering genetic bases of human inherited diseases.

© The Author(s) 2016. Published by Oxford University Press.

DOI: 10.1093/database/baw024

PMCID: PMC4795934

PMID: 26989155 [Indexed for MEDLINE]

307. PeerJ. 2016 Mar 17;4:e1698. doi: 10.7717/peerj.1698. eCollection 2016.

NeisseriaBase: a specialised Neisseria genomic resource and analysis platform.

Zheng W(1), Mutha NV(2), Heydari H(3), Dutta A(2), Siow CC(2), Jakubovics NS(4),

Wee WY(1), Tan SY(1), Ang MY(1), Wong GJ(1), Choo SW(5).

Author information:

(1)Department of Oral Biology and Biomedical Sciences, Faculty of Dentistry,

University of Malaya, Kuala Lumpur, Malaysia; Genome Informatics Research

Laboratory, HIR Building, University of Malaya, Kuala Lumpur, Malaysia. (2)Genome

Informatics Research Laboratory, HIR Building, University of Malaya , Kuala

Lumpur , Malaysia. (3)Genome Informatics Research Laboratory, HIR Building,

University of Malaya, Kuala Lumpur, Malaysia; Computer Science and Engineering

Department, University of NE-Lincoln, Lincoln NE, United States of America.

(4)Centre for Oral Health Research, School of Dental Sciences, Newcastle

University , Newcastle upon Tyne , United Kingdom. (5)Department of Oral Biology

and Biomedical Sciences, Faculty of Dentistry, University of Malaya, Kuala

Lumpur, Malaysia; Genome Informatics Research Laboratory, HIR Building,

University of Malaya, Kuala Lumpur, Malaysia; Genome Solutions Sdn Bhd, Suite 8,

Innovation Incubator UM, Level 5, Research Management & Innovation Complex,

University of Malaya, Kuala Lumpur, Malaysia.

Background. The gram-negative Neisseria is associated with two of the most potent

human epidemic diseases: meningococcal meningitis and gonorrhoea. In both cases,

disease is caused by bacteria colonizing human mucosal membrane surfaces.

Overall, the genus shows great diversity and genetic variation mainly due to its

ability to acquire and incorporate genetic material from a diverse range of

sources through horizontal gene transfer. Although a number of databases exist

for the Neisseria genomes, they are mostly focused on the pathogenic species. In

this present study we present the freely available NeisseriaBase, a database

dedicated to the genus Neisseria encompassing the complete and draft genomes of

15 pathogenic and commensal Neisseria species. Methods. The genomic data were

retrieved from National Center for Biotechnology Information (NCBI) and annotated

using the RAST server which were then stored into the MySQL database. The

protein-coding genes were further analyzed to obtain information such as

calculation of GC content (%), predicted hydrophobicity and molecular weight (Da)

using in-house Perl scripts. The web application was developed following the

secure four-tier web application architecture: (1) client workstation, (2) web

server, (3) application server, and (4) database server. The web interface was

constructed using PHP, JavaScript, jQuery, AJAX and CSS, utilizing the

model-view-controller (MVC) framework. The in-house developed bioinformatics

tools implemented in NeisseraBase were developed using Python, Perl, BioPerl and

R languages. Results. Currently, NeisseriaBase houses 603,500 Coding Sequences

(CDSs), 16,071 RNAs and 13,119 tRNA genes from 227 Neisseria genomes. The

database is equipped with interactive web interfaces. Incorporation of the

JBrowse genome browser in the database enables fast and smooth browsing of

Neisseria genomes. NeisseriaBase includes the standard BLAST program to

facilitate homology searching, and for Virulence Factor Database (VFDB) specific

homology searches, the VFDB BLAST is also incorporated into the database. In

addition, NeisseriaBase is equipped with in-house designed tools such as the

Pairwise Genome Comparison tool (PGC) for comparative genomic analysis and the

Pathogenomics Profiling Tool (PathoProT) for the comparative pathogenomics

analysis of Neisseria strains. Discussion. This user-friendly database not only

provides access to a host of genomic resources on Neisseria but also enables

high-quality comparative genome analysis, which is crucial for the expanding

scientific community interested in Neisseria research. This database is freely

available at http://neisseria.um.edu.my.

DOI: 10.7717/peerj.1698

PMCID: PMC4806638

PMID: 27017950

308. PLoS One. 2016 Mar 16;11(3):e0150965. doi: 10.1371/journal.pone.0150965.

eCollection 2016.

ENTPRISE: An Algorithm for Predicting Human Disease-Associated Amino Acid

Substitutions from Sequence Entropy and Predicted Protein Structures.

Zhou H(1), Gao M(1), Skolnick J(1).

Author information:

(1)Center for the Study of Systems Biology, School of Biology, Georgia Institute

of Technology, Atlanta, Georgia, United States of America.

The advance of next-generation sequencing technologies has made exome sequencing

rapid and relatively inexpensive. A major application of exome sequencing is the

identification of genetic variations likely to cause Mendelian diseases. This

requires processing large amounts of sequence information and therefore

computational approaches that can accurately and efficiently identify the subset

of disease-associated variations are needed. The accuracy and high false positive

rates of existing computational tools leave much room for improvement. Here, we

develop a boosted tree regression machine-learning approach to predict human

disease-associated amino acid variations by utilizing a comprehensive combination

of protein sequence and structure features. On comparing our method, ENTPRISE, to

the state-of-the-art methods SIFT, PolyPhen-2, MUTATIONASSESSOR, MUTATIONTASTER,

FATHMM, ENTPRISE exhibits significant improvement. In particular, on a testing

dataset consisting of only proteins with balanced disease-associated and neutral

variations defined as having the ratio of neutral/disease-associated variations

between 0.3 and 3, the Mathews Correlation Coefficient by ENTPRISE is 0.493 as

compared to 0.432 by PPH2-HumVar, 0.406 by SIFT, 0.403 by MUTATIONASSESSOR, 0.402

by PPH2-HumDiv, 0.305 by MUTATIONTASTER, and 0.181 by FATHMM. ENTPRISE is then

applied to nucleic acid binding proteins in the human proteome.

Disease-associated predictions are shown to be highly correlated with the number

of protein-protein interactions. Both these predictions and the ENTPRISE server

are freely available for academic users as a web service at

http://cssb.biology.gatech.edu/entprise/.

DOI: 10.1371/journal.pone.0150965

PMCID: PMC4794227

PMID: 26982818 [Indexed for MEDLINE]

309. PLoS One. 2016 Mar 16;11(3):e0151323. doi: 10.1371/journal.pone.0151323.

eCollection 2016.

An Approach to Function Annotation for Proteins of Unknown Function (PUFs) in the

Transcriptome of Indian Mulberry.

Dhanyalakshmi KH(1), Naika MB(2), Sajeevan RS(1), Mathew OK(2), Shafi KM(2),

Sowdhamini R(2), N Nataraja K(1).

Author information:

(1)Department of Crop Physiology, University of Agricultural Sciences, GKVK,

Bengaluru, 560065, India. (2)National Centre for Biological Sciences, TIFR, GKVK

campus, Bengaluru, 560065, India.

The modern sequencing technologies are generating large volumes of information at

the transcriptome and genome level. Translation of this information into a

biological meaning is far behind the race due to which a significant portion of

proteins discovered remain as proteins of unknown function (PUFs). Attempts to

uncover the functional significance of PUFs are limited due to lack of easy and

high throughput functional annotation tools. Here, we report an approach to

assign putative functions to PUFs, identified in the transcriptome of mulberry, a

perennial tree commonly cultivated as host of silkworm. We utilized the mulberry

PUFs generated from leaf tissues exposed to drought stress at whole plant level.

A sequence and structure based computational analysis predicted the probable

function of the PUFs. For rapid and easy annotation of PUFs, we developed an

automated pipeline by integrating diverse bioinformatics tools, designated as

PUFs Annotation Server (PUFAS), which also provides a web service API

(Application Programming Interface) for a large-scale analysis up to a genome.

The expression analysis of three selected PUFs annotated by the pipeline revealed

abiotic stress responsiveness of the genes, and hence their potential role in

stress acclimation pathways. The automated pipeline developed here could be

extended to assign functions to PUFs from any organism in general. PUFAS web

server is available at http://caps.ncbs.res.in/pufas/ and the web service is

accessible at http://capservices.ncbs.res.in/help/pufas.

DOI: 10.1371/journal.pone.0151323

PMCID: PMC4794119

PMID: 26982336 [Indexed for MEDLINE]

310. Anal Biochem. 2016 Mar 15;497:60-7. doi: 10.1016/j.ab.2015.12.017. Epub 2015 Dec

31.

pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via

physical-chemical properties.

Liu Z(1), Xiao X(2), Yu DJ(3), Jia J(4), Qiu WR(4), Chou KC(5).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China; School of Computer Science and Engineering, Nanjing University of Science

and Technology, Nanjing 210094, China. (2)Computer Department, Jing-De-Zhen

Ceramic Institute, Jing-De-Zhen 333403, China; Information School, ZheJiang

Textile and Fashion College, NingBo 315211, China; Gordon Life Science Institute,

Boston, MA 02478, USA. Electronic address: xxiao@gordonlifescience.org. (3)School

of Computer Science and Engineering, Nanjing University of Science and

Technology, Nanjing 210094, China. (4)Computer Department, Jing-De-Zhen Ceramic

Institute, Jing-De-Zhen 333403, China. (5)Gordon Life Science Institute, Boston,

MA 02478, USA; Center of Excellence in Genomic Medicine Research (CEGMR), King

Abdulaziz University, Jeddah 21589, Saudi Arabia.

Just like PTM or PTLM (post-translational modification) in proteins, PTCM

(post-transcriptional modification) in RNA plays very important roles in

biological processes. Occurring at adenine (A) with the genetic code motif (GAC),

N(6)-methyldenosine (m(6)A) is one of the most common and abundant PTCMs in RNA

found in viruses and most eukaryotes. Given an uncharacterized RNA sequence

containing many GAC motifs, which of them can be methylated, and which cannot? It

is important for both basic research and drug development to address this

problem. Particularly with the avalanche of RNA sequences generated in the

postgenomic age, it is highly demanded to develop computational methods for

timely identifying the N(6)-methyldenosine sites in RNA. Here we propose a new

predictor called pRNAm-PC, in which RNA sequence samples are expressed by a novel

mode of pseudo dinucleotide composition (PseDNC) whose components were derived

from a physical-chemical matrix via a series of auto-covariance and cross

covariance transformations. It was observed via a rigorous jackknife test that,

in comparison with the existing predictor for the same purpose, pRNAm-PC achieved

remarkably higher success rates in both overall accuracy and stability,

indicating that the new predictor will become a useful high-throughput tool for

identifying methylation sites in RNA, and that the novel approach can also be

used to study many other RNA-related problems and conduct genome analysis. A

user-friendly Web server for pRNAm-PC has been established at

http://www.jci-bioinfo.cn/pRNAm-PC, by which users can easily get their desired

results without needing to go through the mathematical details.

Copyright © 2015 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2015.12.017

PMID: 26748145 [Indexed for MEDLINE]

311. Anal Biochem. 2016 Mar 15;497:48-56. doi: 10.1016/j.ab.2015.12.009. Epub 2015 Dec

23.

iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating

sequence-coupling effects into pseudo components and optimizing imbalanced

training dataset.

Jia J(1), Liu Z(2), Xiao X(3), Liu B(2), Chou KC(4).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China. Electronic address: jjh163yx@163.com. (2)Computer Department, Jing-De-Zhen

Ceramic Institute, Jing-De-Zhen 333403, China. (3)Computer Department,

Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China; Gordon Life Science

Institute, Boston, MA 02478, USA. Electronic address:

xxiao@gordonlifescience.org. (4)Gordon Life Science Institute, Boston, MA 02478,

USA; Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz

University, Jeddah 21589, Saudi Arabia. Electronic address:

kcchou@gordonlifescience.org.

Succinylation is a posttranslational modification (PTM) where a succinyl group is

added to a Lys (K) residue of a protein molecule. Lysine succinylation plays an

important role in orchestrating various biological processes, but it is also

associated with some diseases. Therefore, we are challenged by the following

problem from both basic research and drug development: given an uncharacterized

protein sequence containing many Lys residues, which one of them can be

succinylated, and which one cannot? With the avalanche of protein sequences

generated in the postgenomic age, the answer to the problem has become even more

urgent. Fortunately, the statistical significance experimental data for

succinylated sites in proteins have become available very recently, an

indispensable prerequisite for developing a computational method to address this

problem. By incorporating the sequence-coupling effects into the general pseudo

amino acid composition and using KNNC (K-nearest neighbors cleaning) treatment

and IHTS (inserting hypothetical training samples) treatment to optimize the

training dataset, a predictor called iSuc-PseOpt has been developed. Rigorous

cross-validations indicated that it remarkably outperformed the existing method.

A user-friendly web-server for iSuc-PseOpt has been established at

http://www.jci-bioinfo.cn/iSuc-PseOpt, where users can easily get their desired

results without needing to go through the complicated mathematical equations

involved.

Copyright © 2015 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2015.12.009

PMID: 26723495 [Indexed for MEDLINE]

312. Bioinformatics. 2016 Mar 15;32(6):929-31. doi: 10.1093/bioinformatics/btv681.

Epub 2015 Nov 16.

JSpeciesWS: a web server for prokaryotic species circumscription based on

pairwise genome comparison.

Richter M(1), Rosselló-Móra R(2), Oliver Glöckner F(3), Peplies J(1).

Author information:

(1)Ribocon GmbH, Fahrenheitstraße. 1, 28359 Bremen, Germany. (2)Marine

Microbiology Group, Mediterranean Institute for Advanced Studies, C/ Miquel

Marqués, 21, 07190, Esporles, Spain. (3)Max Planck Institute for Marine

Microbiology, Microbial Genomics and Bioinformatics Research Group, Celsiusstraße

1, 28359 Bremen, Germany and Jacobs University gGmbH Bremen, Department of Life

Sciences and Chemistry, Campus Ring 1, 28759 Bremen, Germany.

JSpecies Web Server (JSpeciesWS) is a user-friendly online service for in silico

calculating the extent of identity between two genomes, a parameter routinely

used in the process of polyphasic microbial species circumscription. The service

measures the average nucleotide identity (ANI) based on BLAST+ (ANIb) and MUMmer

(ANIm), as well as correlation indexes of tetra-nucleotide signatures (Tetra). In

addition, it provides a Tetra Correlation Search function, which allows to

rapidly compare selected genomes against a continuously updated reference

database with currently about 32 000 published whole and draft genome sequences.

For comparison, own genomes can be uploaded and references can be selected from

the JSpeciesWS reference database. The service indicates whether two genomes

share genomic identities above or below the species embracing thresholds, and

serves as a fast way to allocate unknown genomes in the frame of the hitherto

sequenced species.AVAILABILITY AND IMPLEMENTATION: JSpeciesWS is available at

http://jspecies.ribohost.com/jspeciesws

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

CONTACT: mrichter@ribocon.com.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv681

PMID: 26576653

313. Bioinformatics. 2016 Mar 15;32(6):932-6. doi: 10.1093/bioinformatics/btv663. Epub

2015 Nov 14.

KMAD: knowledge-based multiple sequence alignment for intrinsically disordered

proteins.

Lange J(1), Wyrwicz LS(2), Vriend G(3).

Author information:

(1)Laboratory of Bioinformatics and Biostatistics, M. Sklodowska-Curie Memorial

Cancer Center and Institute of Oncology, Warsaw, Poland and Laboratory of

Bioinformatics and Biostatistics, M. Sklodowska-Curie Memorial Cancer Center and

Institute of Oncology, Warsaw, Poland and. (2)Laboratory of Bioinformatics and

Biostatistics, M. Sklodowska-Curie Memorial Cancer Center and Institute of

Oncology, Warsaw, Poland and. (3)CMBI Radboudumc, 6525 GA, Nijmegen, The

Netherlands.

Intrinsically disordered proteins (IDPs) lack tertiary structure and thus differ

from globular proteins in terms of their sequence-structure-function relations.

IDPs have lower sequence conservation, different types of active sites and a

different distribution of functionally important regions, which altogether make

their multiple sequence alignment (MSA) difficult. The KMAD MSA software has been

written specifically for the alignment and annotation of IDPs. It augments the

substitution matrix with knowledge about post-translational modifications,

functional domains and short linear motifs.RESULTS: MSAs produced with KMAD

describe well-conserved features among IDPs, tend to agree well with biological

intuition, and are a good basis for designing new experiments to shed light on

this large, understudied class of proteins.

AVAILABILITY AND IMPLEMENTATION: KMAD web server is accessible at

http://www.cmbi.ru.nl/kmad/ A standalone version is freely available.

CONTACT: vriend@cmbi.ru.nl.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv663

PMCID: PMC4803389

PMID: 26568635

314. Appl Transl Genom. 2016 Mar 12;9:30-2. doi: 10.1016/j.atg.2016.03.002.

eCollection 2016.

SeeHaBITaT: A server on bioinformatics applications for Tospoviruses and other

species.

Sakthivel S(1), Habeeb SK(1).

Author information:

(1)Department of Bioinformatics, School of Bioengineering, SRM University,

Kattankulathur, Chennai 603203, India.

Plant viruses are important limiting factors in agricultural productivity.

Tospovirus is one of the severe plant pathogens, causing damage to economically

important food and ornamental crops worldwide through thrips as vectors. Database

application resources exclusively on this virus would help to design better

control measures, which aren't available. SeeHaBITaT is a unique and exclusive

web based server providing work bench to perform computational research on

tospoviruses and its species. SeeHaBITaT hosts Tospoviruses specific database

Togribase, MOLBIT, SRMBIT and SS with PDB. These applications would be of immense

help to the Tospovirus scientific community. The server could be accessed at

http://bit.srmuniv.ac.in/.

DOI: 10.1016/j.atg.2016.03.002

PMCID: PMC4912379

PMID: 27354938

315. Front Microbiol. 2016 Mar 11;7:283. doi: 10.3389/fmicb.2016.00283. eCollection

2016.

DistAMo: A Web-Based Tool to Characterize DNA-Motif Distribution on Bacterial

Chromosomes.

Sobetzko P(1), Jelonek L(2), Strickert M(1), Han W(1), Goesmann A(2),

Waldminghaus T(1).

Author information:

(1)Chromosome Biology Group, LOEWE Center for Synthetic Microbiology, SYNMIKRO,

Philipps-University Marburg Marburg, Germany. (2)Bioinformatics and Systems

Biology, Faculty of Biology and Chemistry, Justus-Liebig-University Giessen,

Germany.

Short DNA motifs are involved in a multitude of functions such as for example

chromosome segregation, DNA replication or mismatch repair. Distribution of such

motifs is often not random and the specific chromosomal pattern relates to the

respective motif function. Computational approaches which quantitatively assess

such chromosomal motif patterns are necessary. Here we present a new computer

tool DistAMo (Distribution Analysis of DNA Motifs). The algorithm uses codon

redundancy to calculate the relative abundance of short DNA motifs from single

genes to entire chromosomes. Comparative genomics analyses of the GATC-motif

distribution in γ-proteobacterial genomes using DistAMo revealed that (i) genes

beside the replication origin are enriched in GATCs, (ii) genome-wide GATC

distribution follows a distinct pattern, and (iii) genes involved in DNA

replication and repair are enriched in GATCs. These features are specific for

bacterial chromosomes encoding a Dam methyltransferase. The new software is

available as a stand-alone or as an easy-to-use web-based server version at

http://www.computational.bio.uni-giessen.de/distamo.

DOI: 10.3389/fmicb.2016.00283

PMCID: PMC4786541

PMID: 27014208

316. PLoS Comput Biol. 2016 Mar 10;12(3):e1004794. doi: 10.1371/journal.pcbi.1004794.

eCollection 2016.

SMOG 2: A Versatile Software Package for Generating Structure-Based Models.

Noel JK(1,)(2), Levi M(3), Raghunathan M(1), Lammert H(1), Hayes RL(1), Onuchic

JN(1), Whitford PC(3).

Author information:

(1)Center for Theoretical Biological Physics, Rice University, Houston, Texas,

United States of America. (2)Kristallografie, Max Delbrück Center for Molecular

Medicine, Berlin, Germany. (3)Department of Physics, Northeastern University,

Boston, Massachusetts, United States of America.

Molecular dynamics simulations with coarse-grained or simplified Hamiltonians

have proven to be an effective means of capturing the functionally important

long-time and large-length scale motions of proteins and RNAs. Originally

developed in the context of protein folding, structure-based models (SBMs) have

since been extended to probe a diverse range of biomolecular processes, spanning

from protein and RNA folding to functional transitions in molecular machines. The

hallmark feature of a structure-based model is that part, or all, of the

potential energy function is defined by a known structure. Within this general

class of models, there exist many possible variations in resolution and energetic

composition. SMOG 2 is a downloadable software package that reads user-designated

structural information and user-defined energy definitions, in order to produce

the files necessary to use SBMs with high performance molecular dynamics

packages: GROMACS and NAMD. SMOG 2 is bundled with XML-formatted template files

that define commonly used SBMs, and it can process template files that are

altered according to the needs of each user. This computational infrastructure

also allows for experimental or bioinformatics-derived restraints or novel

structural features to be included, e.g. novel ligands, prosthetic groups and

post-translational/transcriptional modifications. The code and user guide can be

downloaded at http://smog-server.org/smog2.

DOI: 10.1371/journal.pcbi.1004794

PMCID: PMC4786265

PMID: 26963394 [Indexed for MEDLINE]

317. Biotechnol Biofuels. 2016 Mar 8;9:54. doi: 10.1186/s13068-016-0471-8. eCollection

2016.

Dissection of early transcriptional responses to water stress in Arundo donax L.

by unigene-based RNA-seq.

Fu Y(1), Poli M(2), Sablok G(3), Wang B(4), Liang Y(5), La Porta N(6), Velikova

V(7), Loreto F(8), Li M(3), Varotto C(3).

Author information:

(1)Department of Biodiversity and Molecular Ecology, Research and Innovation

Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 San Michele all'Adige,

Trento Italy ; Dipartimento di Biotecnologie, Università degli Studi di Verona,

Verona, Italy. (2)Department of Biodiversity and Molecular Ecology, Research and

Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 San Michele

all'Adige, Trento Italy ; Dipartimento di Scienze Agrarie, Università di Bologna,

Bologna, Italy. (3)Department of Biodiversity and Molecular Ecology, Research and

Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 San Michele

all'Adige, Trento Italy. (4)Department of Biodiversity and Molecular Ecology,

Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 San

Michele all'Adige, Trento Italy ; Centro di Biologia Integrata (CIBIO),

University of Trento, Trento, Italy. (5)College of Computer Science and

Technology, Jilin University, Changchun, China. (6)Department of Biodiversity and

Molecular Ecology, Research and Innovation Centre, Fondazione Edmund Mach, Via E.

Mach 1, 38010 San Michele all'Adige, Trento Italy ; MOUNTFOR Project Centre,

European Forest Institute, Via E. Mach 1, 38010 San Michele all'Adige, Trento

Italy. (7)Department of Biodiversity and Molecular Ecology, Research and

Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 San Michele

all'Adige, Trento Italy ; Institute of Plant Physiology and Genetics, Bulgarian

Academy of Sciences, Sofia, Bulgaria. (8)The National Research Council of Italy

(CNR), Department of Biology, Agriculture and Food Sciences, Rome, Italy.

BACKGROUND: Arundo donax L. (Poaceae) is considered one of the most promising

energy crops in the Mediterranean region because of its high biomass yield and

low input requirements, but to date no information on its transcriptional

responses to water stress is available.

RESULTS: We obtained by Illumina-based RNA-seq the whole root and shoot

transcriptomes of young A. donax plants subjected to osmotic/water stress with 10

and 20 % polyethylene glycol (PEG; 3 biological replicates/organ/condition

corresponding to 18 RNA-Seq libraries), and identified a total of 3034

differentially expressed genes. Blast-based mining of stress-related genes

indicated the higher responsivity of roots compared to shoots at the early stages

of water stress especially under the milder PEG treatment, with a majority of

genes responsive to salt, oxidative, and dehydration stress. Analysis of gene

ontology terms underlined the qualitatively different responses between root and

shoot tissues. Among the most significantly enriched metabolic pathways

identified using a Fisher's exact test with FDR correction, a crucial role was

played in both shoots and roots by genes involved in the signaling cascade of

abscisic acid. We further identified relatively large organ-specific differences

in the patterns of drought-related transcription factor AP2-EREBP, AUX/IAA, MYB,

bZIP, C2H2, and GRAS families, which may underlie the transcriptional

reprogramming differences between organs. Through comparative analyses with major

Poaceae species based on Blast, we finally identified a set of 53 orthologs that

can be considered as a core of evolutionary conserved genes important to mediate

water stress responses in the family.

CONCLUSIONS: This study provides the first characterization of A. donax

transcriptome in response to water stress, thus shedding novel light at the

molecular level on the mechanisms of stress response and adaptation in this

emerging bioenergy species. The inventory of early-responsive genes to water

stress identified could constitute useful markers of the physiological status of

A. donax and be a basis for the improvement of its productivity under water

limitation. The full water-stressed A. donax transcriptome is available for

Blast-based homology searches through a dedicated web server

(http://ecogenomics.fmach.it/arundo/).

DOI: 10.1186/s13068-016-0471-8

PMCID: PMC4782572

PMID: 26958077

318. BMC Bioinformatics. 2016 Mar 8;17:119. doi: 10.1186/s12859-016-0975-z.

CoeViz: a web-based tool for coevolution analysis of protein residues.

Baker FN(1,)(2), Porollo A(3,)(4).

Author information:

(1)Department of Electrical Engineering and Computing Systems, University of

Cincinnati, 2901 Woodside Drive, Cincinnati, OH, 45221, USA. bakerfn@mail.uc.edu.

(2)Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital

Medical Center, 3333 Burnet Avenue, Cincinnati, OH, 45229, USA.

bakerfn@mail.uc.edu. (3)Center for Autoimmune Genomics and Etiology, Cincinnati

Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH, 45229,

USA. alexey.porollo@cchmc.org. (4)Division of Biomedical Informatics, Cincinnati

Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH, 45229,

USA. alexey.porollo@cchmc.org.

BACKGROUND: Proteins generally perform their function in a folded state. Residues

forming an active site, whether it is a catalytic center or interaction

interface, are frequently distant in a protein sequence. Hence, traditional

sequence-based prediction methods focusing on a single residue (or a short window

of residues) at a time may have difficulties in identifying and clustering the

residues constituting a functional site, especially when a protein has multiple

functions. Evolutionary information encoded in multiple sequence alignments is

known to greatly improve sequence-based predictions. Identification of coevolving

residues further advances the protein structure and function annotation by

revealing cooperative pairs and higher order groupings of residues.

RESULTS: We present a new web-based tool (CoeViz) that provides a versatile

analysis and visualization of pairwise coevolution of amino acid residues. The

tool computes three covariance metrics: mutual information, chi-square statistic,

Pearson correlation, and one conservation metric: joint Shannon entropy.

Implemented adjustments of covariance scores include phylogeny correction,

corrections for sequence dissimilarity and alignment gaps, and the average

product correction. Visualization of residue relationships is enhanced by

hierarchical cluster trees, heat maps, circular diagrams, and the residue

highlighting in protein sequence and 3D structure. Unlike other existing tools,

CoeViz is not limited to analyzing conserved domains or protein families and can

process long, unstructured and multi-domain proteins thousands of residues long.

Two examples are provided to illustrate the use of the tool for identification of

residues (1) involved in enzymatic function, (2) forming short linear functional

motifs, and (3) constituting a structural domain.

CONCLUSIONS: CoeViz represents a practical resource for a quick sequence-based

protein annotation for molecular biologists, e.g., for identifying putative

functional clusters of residues and structural domains. CoeViz also can serve

computational biologists as a resource of coevolution matrices, e.g., for

developing machine learning-based prediction models. The presented tool is

integrated in the POLYVIEW-2D server (http://polyview.cchmc.org/) and available

from resulting pages of POLYVIEW-2D.

DOI: 10.1186/s12859-016-0975-z

PMCID: PMC4782369

PMID: 26956673 [Indexed for MEDLINE]

319. Sci Rep. 2016 Mar 8;6:22843. doi: 10.1038/srep22843.

A Web Server and Mobile App for Computing Hemolytic Potency of Peptides.

Chaudhary K(1), Kumar R(2), Singh S(1), Tuknait A(1), Gautam A(1), Mathur D(1),

Anand P(1), Varshney GC(3), Raghava GP(1).

Author information:

(1)Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39A,

Chandigarh, India. (2)CSIR-Central Scientific Instruments Organisation, Sector

30C, Chandigarh, India. (3)Cell Biology and Immunology Division, CSIR-Institute

of Microbial Technology, Chandigarh, India.

Numerous therapeutic peptides do not enter the clinical trials just because of

their high hemolytic activity. Recently, we developed a database, Hemolytik, for

maintaining experimentally validated hemolytic and non-hemolytic peptides. The

present study describes a web server and mobile app developed for predicting, and

screening of peptides having hemolytic potency. Firstly, we generated a dataset

HemoPI-1 that contains 552 hemolytic peptides extracted from Hemolytik database

and 552 random non-hemolytic peptides (from Swiss-Prot). The sequence analysis of

these peptides revealed that certain residues (e.g., L, K, F, W) and motifs

(e.g., "FKK", "LKL", "KKLL", "KWK", "VLK", "CYCR", "CRR", "RFC", "RRR", "LKKL")

are more abundant in hemolytic peptides. Therefore, we developed models for

discriminating hemolytic and non-hemolytic peptides using various machine

learning techniques and achieved more than 95% accuracy. We also developed models

for discriminating peptides having high and low hemolytic potential on different

datasets called HemoPI-2 and HemoPI-3. In order to serve the scientific

community, we developed a web server, mobile app and JAVA-based standalone

software (http://crdd.osdd.net/raghava/hemopi/).

DOI: 10.1038/srep22843

PMCID: PMC4782144

PMID: 26953092 [Indexed for MEDLINE]

320. Behav Res Methods. 2016 Mar;48(1):123-37. doi: 10.3758/s13428-014-0560-1.

LSE-Sign: A lexical database for Spanish Sign Language.

Gutierrez-Sigut E(1), Costello B(2), Baus C(3), Carreiras M(4,)(5,)(6).

Author information:

(1)Deafness, Cognition and Language Research Centre, University College London,

London, UK. (2)BCBL. Basque Center on Cognition, Brain and Language, Paseo

Mikeletegi 69-2, 20009, Donostia, Spain. b.costello@bcbl.eu. (3)Laboratoire de

Psychologie Cognitive, CNRS and Université d'Aix-Marseille, Marseille, France.

(4)BCBL. Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69-2,

20009, Donostia, Spain. (5)Ikerbasque, Basque Foundation for Science, Bilbao,

Spain. (6)Departamento de Lengua Vasca y Comunicación, Universidad del País Vasco

UPV/EHU, Donostia, Spain.

The LSE-Sign database is a free online tool for selecting Spanish Sign Language

stimulus materials to be used in experiments. It contains 2,400 individual signs

taken from a recent standardized LSE dictionary, and a further 2,700 related

nonsigns. Each entry is coded for a wide range of grammatical, phonological, and

articulatory information, including handshape, location, movement, and non-manual

elements. The database is accessible via a graphically based search facility

which is highly flexible both in terms of the search options available and the

way the results are displayed. LSE-Sign is available at the following website:

http://www.bcbl.eu/databases/lse/.

DOI: 10.3758/s13428-014-0560-1

PMID: 25630312 [Indexed for MEDLINE]

321. Bioinformatics. 2016 Mar 1;32(5):779-81. doi: 10.1093/bioinformatics/btv645. Epub

2015 Nov 2.

No Promoter Left Behind (NPLB): learn de novo promoter architectures from

genome-wide transcription start sites.

Mitra S(1), Narlikar L(1).

Author information:

(1)Chemical Engineering Division, CSIR-National Chemical Laboratory, Pune 411008,

India.

Promoters have diverse regulatory architectures and thus activate genes

differently. For example, some have a TATA-box, many others do not. Even the ones

with it can differ in its position relative to the transcription start site

(TSS). No Promoter Left Behind (NPLB) is an efficient, organism-independent

method for characterizing such diverse architectures directly from experimentally

identified genome-wide TSSs, without relying on known promoter elements. As a

test case, we show its application in identifying novel architectures in the fly

genome.AVAILABILITY AND IMPLEMENTATION: Web-server at http://nplb.ncl.res.in

Standalone also at https://github.com/computationalBiology/NPLB/ (Mac OSX/Linux).

CONTACT: l.narlikar@ncl.res.in

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv645

PMCID: PMC4795619

PMID: 26530723

322. Bioinformatics. 2016 Mar 1;32(5):776-8. doi: 10.1093/bioinformatics/btv640. Epub

2015 Oct 30.

Riboswitch Scanner: an efficient pHMM-based web-server to detect riboswitches in

genomic sequences.

Mukherjee S(1), Sengupta S(1).

Author information:

(1)Department of Physical Sciences, Indian Institute of Science Education and

Research Kolkata, Mohanpur-741246, West Bengal, India.

Riboswitches are non-coding RNA located in the 5' untranslated regions where they

bind a target metabolite used to specify the riboswitch class and control the

expression of associated genes. Accurate identification of riboswitches is the

first step towards understanding their regulatory and functional roles in the

cell. In this article, we describe a new web application named Riboswitch Scanner

which provides an automated pipeline for pHMM-based detection of riboswitches in

partial as well as complete genomic sequences rapidly, with high sensitivity and

specificity.AVAILABILITY AND IMPLEMENTATION: Riboswitch Scanner can be freely

accessed on the web at http://service.iiserkol.ac.in/∼riboscan/

CONTACT: mukherjee.sumit89@gmail.com

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv640

PMID: 26519506

323. Mol Biosyst. 2016 Mar;12(3):786-95. doi: 10.1039/c5mb00853k. Epub 2016 Jan 7.

SuccinSite: a computational tool for the prediction of protein succinylation

sites by exploiting the amino acid patterns and properties.

Hasan MM(1), Yang S, Zhou Y, Mollah MN.

Author information:

(1)State Key Laboratory of Agrobiotechnology, College of Biological Sciences,

China Agricultural University, Beijing, 100193, China. mehedicau@hotmail.com.

Lysine succinylation is an emerging protein post-translational modification,

which plays an important role in regulating the cellular processes in both

eukaryotic and prokaryotic cells. However, the succinylation modification site is

particularly difficult to detect because the experimental technologies used are

often time-consuming and costly. Thus, an accurate computational method for

predicting succinylation sites may help researchers towards designing their

experiments and to understand the molecular mechanism of succinylation. In this

study, a novel computational tool termed SuccinSite has been developed to predict

protein succinylation sites by incorporating three sequence encodings, i.e.,

k-spaced amino acid pairs, binary and amino acid index properties. Then, the

random forest classifier was trained with these encodings to build the predictor.

The SuccinSite predictor achieves an AUC score of 0.802 in the 5-fold

cross-validation set and performs significantly better than existing predictors

on a comprehensive independent test set. Furthermore, informative features and

predominant rules (i.e. feature combinations) were extracted from the trained

random forest model for an improved interpretation of the predictor. Finally, we

also compiled a database covering 4411 experimentally verified succinylation

proteins with 12 456 lysine succinylation sites. Taken together, these results

suggest that SuccinSite would be a helpful computational resource for

succinylation sites prediction. The web-server, datasets, source code and

database are freely available at http://systbio.cau.edu.cn/SuccinSite/.

DOI: 10.1039/c5mb00853k

PMID: 26739209 [Indexed for MEDLINE]

324. Mol Genet Metab. 2016 Mar;117(3):322-7. doi: 10.1016/j.ymgme.2015.12.007. Epub

2015 Dec 23.

Metabolic Diet App Suite for inborn errors of amino acid metabolism.

Ho G(1), Ueda K(1), Houben RF(2), Joa J(2), Giezen A(1), Cheng B(1), van

Karnebeek CD(3).

Author information:

(1)Division of Biochemical Diseases, BC Children's Hospital, University of

British Columbia, Vancouver, Canada. (2)Health2Media, Vancouver, Canada.

(3)Division of Biochemical Diseases, BC Children's Hospital, University of

British Columbia, Vancouver, Canada; Department of Pediatrics, Centre for

Molecular Medicine & Therapeutics, Child and Family Research Institute,

University of British Columbia, Vancouver, Canada. Electronic address:

cvankarnebeek@cw.bc.ca.

BACKGROUND: An increasing number of rare inborn errors of metabolism (IEMs) are

amenable to targeted metabolic nutrition therapy. Daily adherence is important to

attain metabolic control and prevent organ damage. This is challenging however,

given the lack of information of disorder specific nutrient content of foods, the

limited availability and cost of specialty products as well as difficulties in

reliable calculation and tracking of dietary intake and targets.

OBJECTIVES: To develop apps for all inborn errors of amino acid metabolism for

which the mainstay of treatment is a medical diet, and obtain patient and family

feedback throughout the process to incorporate this into subsequent versions.

METHODS & RESULTS: The Metabolic Diet App Suite was created with input from

health care professionals as a free, user-friendly, online tool for both mobile

devices and desktop computers (http://www.metabolicdietapp.org) for 15 different

IEMs. General information is provided for each IEM with links to useful online

resources. Nutrient information is based on the MetabolicPro™, a North American

food database compiled by the Genetic Metabolic Dietitians International (GMDI)

Technology committee. After user registration, a personalized dashboard and

management plan including specific nutrient goals are created. Each Diet App has

a user-friendly interface and the functions include: nutrient intake counts,

adding your own foods and homemade recipes and, managing a daily food diary.

Patient and family feedback was overall positive and specific suggestions were

used to further improve the App Suite.

DISCUSSION: The Metabolic Diet App Suite aids individuals affected by IEMs to

track and plan their meals. Future research should evaluate its impact on patient

adherence, metabolic control, quality of life and health-related outcomes. The

Suite will be updated and expanded to Apps for other categories of IEMs. Finally,

this Suite is a support tool only, and does not replace medical/metabolic

nutrition professional advice.

Copyright © 2015 The Authors. Published by Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ymgme.2015.12.007

PMID: 26748688 [Indexed for MEDLINE]

325. Stroke. 2016 Mar;47(3):882-5. doi: 10.1161/STROKEAHA.115.011930. Epub 2016 Jan 7.

Online Tool to Improve Stratification of Adverse Events in Stroke Clinical

Trials.

Hesse K(1), MacIsaac RL(1), Abdul-Rahim AH(2), Lyden PD(1), Bluhmki E(1), Lees

KR(1); VISTA Collaborators.

Collaborators: Alexandrov A, Bath PM, Bornstein N, Claesson L, Davis SM, Donnan

G, Diener HC, Fisher M, Gregson B, Grotta J, Hacke W, Hennerici MG, Hommel M,

Kaste M, Marler J, Muir K, Sacco R, Shuaib A, Teal P, Wahlgren NG, Warach S,

Weimar C.

Author information:

(1)From the Medical School (K.H.), Institute of Cardiovascular and Medical

Sciences (A.H.A.-R, K.R.L., R.L.M.), University of Glasgow, Glasgow, United

Kingdom; Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, CA

(P.D.L.); and Department of Statistics, Boehringer Ingelheim, Bieberach, Germany

(E.B.). (2)From the Medical School (K.H.), Institute of Cardiovascular and

Medical Sciences (A.H.A.-R, K.R.L., R.L.M.), University of Glasgow, Glasgow,

United Kingdom; Department of Neurology, Cedars-Sinai Medical Center, Los

Angeles, CA (P.D.L.); and Department of Statistics, Boehringer Ingelheim,

Bieberach, Germany (E.B.). Azmil.Abdul-Rahim@glasgow.ac.uk.

BACKGROUND AND PURPOSE: Knowing characteristic adverse events (AEs) and their

incidence among patients participating in acute stroke trials may assist

interpretation of future studies. We aimed to develop an online tool to inform

stroke trial safety.

METHODS: We identified relevant AEs from patients within the Virtual

International Stroke Trials Archive (VISTA), using receiver operating

characteristic principles. We modeled their incidence on patient age, baseline

National Institutes of Health Stroke Scale, and comorbidities using binary

logistic regression. Models with an R(2) >5% were deemed powerful enough to

predict expected AE incidences and were included. The calculator was developed

using programs R and Visual Studios.

RESULTS: Forty-eight of the most common AEs were identified and incorporated into

the IschAEmic Stroke Calculator. The calculator, publicly available at

http://www.vistacollaboration.org calculates the expected incidence of AEs or

groups of AEs in a trial cohort and where possible compares them with the

observed incidence.

CONCLUSIONS: The IschAEmic Stroke Calculator is an open access resource to

support safety interpretation within acute stroke trials. Prediction of AEs with

higher likelihood of occurrence may direct preventive clinical measures.

© 2016 American Heart Association, Inc.

DOI: 10.1161/STROKEAHA.115.011930

PMID: 26742798 [Indexed for MEDLINE]

326. Zh Mikrobiol Epidemiol Immunobiol. 2016 Mar-Apr;(2):16-23.

[CHARACTERISTICS OF BIOLOGICAL AND MOLECULAR-GENETIC PROPERTIES OF LACTOBACILLUS

FERMENTUM 90 TC-4 PROBIOTIC STRAIN].

[Article in Russian]

Tochilina AG, Belova IV, Solovieva IV, Gorlova IS, Ivanova TP, Zhirnov VA.

AIM: Confirmation of taxonomic position of Lactobacillus fermentum 90 TC-4 strain

using phenotypic (classic microbiological, MALDI TOF mass-spectrometry) and

genetic (16S rRNA gene segment sequencing and full genome sequencing) methods.

MATERIALS AND METHODS: Object of the study--Lactobacillus fermentum 90 TC-4

strains from various collections. Mass-spectrometric analysis was carried out

using Autoflex MALDI TOF mass-spectrometer (Bruker Daltonics, Germany), study of

biochemical properties of the strain was carried out using API 50 CHL strips

(Biomerueux, France), "DNA-sorb B" kitwas used for isolation ofgenome DNA (CRIE,

Moscow). Sequencing of the accumulated fragments of 16S rRNA gene was carried out

using GenomeLab GeXP sequencing (Beckman Coulter, USA), full genome sequencing

was carried out in MiSeq platform (Illumina). Assembly of genome and

bioinformation analysis was carried out using BLAST program

(www.blast.ncbi.nlm.nih.gov/blast.cgi), "CLC Bio Assembly" and genome server RAST

(rast.nmpdr.org).

RESULTS: L. fermentum 90 TC-4 strain was established to be contaminated by L.

plantarum culture in a series of cases. As a result of identification of a pure

culture of L. fermentum 90 TC-4 strain using a specter of high-technology

methods, membership of the strain in L. fer- mentum species has been proven.

CONCLUSION: Taxonomic status of L. fermentum 90 TC-4 strain was confirmed.

PMID: 27228666 [Indexed for MEDLINE]

327. PLoS One. 2016 Feb 26;11(2):e0149621. doi: 10.1371/journal.pone.0149621.

eCollection 2016.

The Implicitome: A Resource for Rationalizing Gene-Disease Associations.

Hettne KM(1), Thompson M(1), van Haagen HH(1), van der Horst E(1), Kaliyaperumal

R(1), Mina E(1), Tatum Z(1), Laros JF(1), van Mulligen EM(1,)(2), Schuemie M(2),

Aten E(1), Li TS(3), Bruskiewich R(4), Good BM(3), Su AI(3), Kors JA(2), den

Dunnen J(1), van Ommen GJ(1), Roos M(1), 't Hoen PA(1), Mons B(1,)(5), Schultes

EA(1,)(6).

Author information:

(1)Department of Human Genetics, Leiden University Medical Center, Leiden, The

Netherlands. (2)Department of Medical Informatics, Erasmus University Medical

Center Rotterdam, Rotterdam, The Netherlands. (3)Department of Molecular and

Experimental Medicine, The Scripps Research Institute, La Jolla, CA, United

States of America. (4)STAR Informatics / Delphinai Corporation, Port Moody, BC,

Canada. (5)Dutch Techcentre for Life Sciences, Utrecht, The Netherlands.

(6)Leiden Institute for Advanced Computer Science, Leiden, The Netherlands.

High-throughput experimental methods such as medical sequencing and genome-wide

association studies (GWAS) identify increasingly large numbers of potential

relations between genetic variants and diseases. Both biological complexity

(millions of potential gene-disease associations) and the accelerating rate of

data production necessitate computational approaches to prioritize and

rationalize potential gene-disease relations. Here, we use concept profile

technology to expose from the biomedical literature both explicitly stated

gene-disease relations (the explicitome) and a much larger set of implied

gene-disease associations (the implicitome). Implicit relations are largely

unknown to, or are even unintended by the original authors, but they vastly

extend the reach of existing biomedical knowledge for identification and

interpretation of gene-disease associations. The implicitome can be used in

conjunction with experimental data resources to rationalize both known and novel

associations. We demonstrate the usefulness of the implicitome by rationalizing

known and novel gene-disease associations, including those from GWAS. To

facilitate the re-use of implicit gene-disease associations, we publish our data

in compliance with FAIR Data Publishing recommendations

[https://www.force11.org/group/fairgroup] using nanopublications. An online tool

(http://knowledge.bio) is available to explore established and potential

gene-disease associations in the context of other biomedical relations.

DOI: 10.1371/journal.pone.0149621

PMCID: PMC4769089

PMID: 26919047 [Indexed for MEDLINE]

328. Sci Rep. 2016 Feb 25;6:21839. doi: 10.1038/srep21839.

dPABBs: A Novel in silico Approach for Predicting and Designing Anti-biofilm

Peptides.

Sharma A(1,)(2), Gupta P(1,)(3), Kumar R(1,)(2), Bhardwaj A(1,)(2).

Author information:

(1)Open Source Drug Discovery (OSDD) Unit, Council of Scientific and Industrial

Research (CSIR), New Delhi, India. (2)Academy of Scientific and Innovative

Research (AcSIR), CSIR-OSDD Unit, CSIR-HQ, New Delhi, India. (3)Bhaskaracharya

College of Applied Sciences, University of Delhi, New Delhi, India.

Increasingly, biofilms are being recognised for their causative role in

persistent infections (like cystic fibrosis, otitis media, diabetic foot ulcers)

and nosocomial diseases (biofilm-infected vascular catheters, implants and

prosthetics). Given the clinical relevance of biofilms and their recalcitrance to

conventional antibiotics, it is imperative that alternative therapeutics are

proactively sought. We have developed dPABBs, a web server that facilitates the

prediction and design of anti-biofilm peptides. The six SVM and Weka models

implemented on dPABBs were observed to identify anti-biofilm peptides on the

basis of their whole amino acid composition, selected residue features and the

positional preference of the residues (maximum accuracy, sensitivity, specificity

and MCC of 95.24%, 92.50%, 97.73% and 0.91, respectively, on the training

datasets). On the N-terminus, it was seen that either of the cationic polar

residues, R and K, is present at all five positions in case of the anti-biofilm

peptides, whereas in the QS peptides, the uncharged polar residue S is

preponderant at the first (also anionic polar residues D, E), third and fifth

positions. Positive predictions were also obtained for 29 FDA-approved peptide

drugs and ten antimicrobial peptides in clinical development, indicating at their

possible repurposing for anti-biofilm therapy. dPABBs is freely accessible on:

http://ab-openlab.csir.res.in/abp/antibiofilm/.

DOI: 10.1038/srep21839

PMCID: PMC4766436

PMID: 26912180 [Indexed for MEDLINE]

329. BMC Bioinformatics. 2016 Feb 24;17:97. doi: 10.1186/s12859-016-0940-x.

Sparse regressions for predicting and interpreting subcellular localization of

multi-label proteins.

Wan S(1), Mak MW(2), Kung SY(3).

Author information:

(1)Department of Electronic and Information Engineering, The Hong Kong

Polytechnic University, Hong Kong, SAR, China. shibiao.wan@connect.polyu.hk.

(2)Department of Electronic and Information Engineering, The Hong Kong

Polytechnic University, Hong Kong, SAR, China. enmwmak@polyu.edu.hk.

(3)Department of Electrical Engineering, Princeton University, New Jersey, USA.

kung@princeton.edu.

BACKGROUND: Predicting protein subcellular localization is indispensable for

inferring protein functions. Recent studies have been focusing on predicting not

only single-location proteins, but also multi-location proteins. Almost all of

the high performing predictors proposed recently use gene ontology (GO) terms to

construct feature vectors for classification. Despite their high performance,

their prediction decisions are difficult to interpret because of the large number

of GO terms involved.

RESULTS: This paper proposes using sparse regressions to exploit GO information

for both predicting and interpreting subcellular localization of single- and

multi-location proteins. Specifically, we compared two multi-label sparse

regression algorithms, namely multi-label LASSO (mLASSO) and multi-label elastic

net (mEN), for large-scale predictions of protein subcellular localization. Both

algorithms can yield sparse and interpretable solutions. By using the one-vs-rest

strategy, mLASSO and mEN identified 87 and 429 out of more than 8,000 GO terms,

respectively, which play essential roles in determining subcellular localization.

More interestingly, many of the GO terms selected by mEN are from the biological

process and molecular function categories, suggesting that the GO terms of these

categories also play vital roles in the prediction. With these essential GO

terms, not only where a protein locates can be decided, but also why it resides

there can be revealed.

CONCLUSIONS: Experimental results show that the output of both mEN and mLASSO are

interpretable and they perform significantly better than existing

state-of-the-art predictors. Moreover, mEN selects more features and performs

better than mLASSO on a stringent human benchmark dataset. For readers'

convenience, an online server called SpaPredictor for both mLASSO and mEN is

available at http://bioinfo.eie.polyu.edu.hk/SpaPredictorServer/.

DOI: 10.1186/s12859-016-0940-x

PMCID: PMC4765148

PMID: 26911432 [Indexed for MEDLINE]

330. Sci Rep. 2016 Feb 24;6:21383. doi: 10.1038/srep21383.

Crysalis: an integrated server for computational analysis and design of protein

crystallization.

Wang H(1), Feng L(1), Zhang Z(2), Webb GI(3), Lin D(1), Song J(3,)(4,)(5).

Author information:

(1)Department of Chemical Biology, College of Chemistry and Chemical Engineering,

Xiamen University, Xiamen 361005, China. (2)State Key Laboratory of

Agrobiotechnology, College of Biological Sciences, China Agricultural University,

Beijing 100193, China. (3)Monash Centre for Data Science, Faculty of Information

Technology, Monash University, Melbourne, VIC 3800, Australia. (4)Department of

Microbiology, School of Biomedical Sciences, Monash University, Melbourne, VIC

3800, Australia. (5)National Engineering Laboratory for Industrial Enzymes,

Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences,

Tianjin 300308, China.

The failure of multi-step experimental procedures to yield diffraction-quality

crystals is a major bottleneck in protein structure determination. Accordingly,

several bioinformatics methods have been successfully developed and employed to

select crystallizable proteins. Unfortunately, the majority of existing in silico

methods only allow the prediction of crystallization propensity, seldom enabling

computational design of protein mutants that can be targeted for enhancing

protein crystallizability. Here, we present Crysalis, an integrated

crystallization analysis tool that builds on support-vector regression (SVR)

models to facilitate computational protein crystallization prediction, analysis,

and design. More specifically, the functionality of this new tool includes: (1)

rapid selection of target crystallizable proteins at the proteome level, (2)

identification of site non-optimality for protein crystallization and systematic

analysis of all potential single-point mutations that might enhance protein

crystallization propensity, and (3) annotation of target protein based on

predicted structural properties. We applied the design mode of Crysalis to

identify site non-optimality for protein crystallization on a proteome-scale,

focusing on proteins currently classified as non-crystallizable. Our results

revealed that site non-optimality is based on biases related to residues,

predicted structures, physicochemical properties, and sequence loci, which

provides in-depth understanding of the features influencing protein

crystallization. Crysalis is freely available at

http://nmrcen.xmu.edu.cn/crysalis/.

DOI: 10.1038/srep21383

PMCID: PMC4764925

PMID: 26906024 [Indexed for MEDLINE]

331. Vet Microbiol. 2016 Feb 23. pii: S0378-1135(16)30042-6. doi:

10.1016/j.vetmic.2016.02.017. [Epub ahead of print]

Identification of Shiga toxin-producing (STEC) and enteropathogenic (EPEC)

Escherichia coli in diarrhoeic calves and comparative genomics of O5 bovine and

human STEC.

Fakih I(1), Thiry D(1), Duprez JN(1), Saulmont M(2), Iguchi A(3), Piérard D(4),

Jouant L(1), Daube G(5), Ogura Y(6), Hayashi T(6), Taminiau B(5), Mainil JG(7).

Author information:

(1)Bacteriology, Infectious Disease Department, Institute for Fundamental and

Applied Research in Animal Health (FARAH) and Faculty of Veterinary Medicine,

University of Liège, 4000 Liège, Belgium. (2)Association Régionale de Santé et

d'Identification Animale (ARSIA), 5590 Ciney, Belgium. (3)Department of Animal

and Grassland Sciences, Faculty of Agriculture, University of Miyazaki, Miyazaki

889-1692, Japan. (4)Dienst Microbiologie en Ziekenhuishygiëne, Universitair

Ziekenhuis Brussel, Vrije Universiteit Brussel, 1090 Brussels, Belgium.

(5)Microbiology, Food Science Department, Institute for Fundamental and Applied

Research in Animal Health (FARAH) and Faculty of Veterinary Medicine, University

of Liège, 4000 Liège, Belgium. (6)Department of Bacteriology, Faculty of Medical

Sciences, Kyushu University, Fukuoka 812-8582, Japan. (7)Bacteriology, Infectious

Disease Department, Institute for Fundamental and Applied Research in Animal

Health (FARAH) and Faculty of Veterinary Medicine, University of Liège, 4000

Liège, Belgium. Electronic address: jg.mainil@ulg.ac.be.

Escherichia coli producing Shiga toxins (Stx) and the attaching-effacing (AE)

lesion (AE-STEC) are responsible for (bloody) diarrhoea in humans and calves

while the enteropathogenic E. coli (EPEC) producing the AE lesion only cause

non-bloody diarrhoea in all mammals. The purpose of this study was (i) to

identify the pathotypes of enterohaemolysin-producing E. coli isolated between

2009 and 2013 on EHLY agar from less than 2 month-old diarrhoeic calves with a

triplex PCR targeting the stx1, stx2, eae virulence genes; (ii) to serotype the

positive isolates with PCR targeting the genes coding for ten most frequent and

pathogenic human and calf STEC O serogroups; and (iii) to compare the MLSTypes

and virulotypes of calf and human O5 AE-STEC after Whole Genome Sequencing using

two server databases (www.genomicepidemiology.org). Of 233 isolates, 206 were

triplex PCR-positive: 119 AE-STEC (58%), 78 EPEC (38%) and 9 STEC (4%); and the

stx1+eae+ AE-STEC (49.5%) were the most frequent. Of them, 120 isolates (84% of

AE-STEC, 23% of EPEC, 22% of STEC) tested positive with one O serogroup PCR: 57

for O26 (47.5%), 36 for O111 (30%), 10 for O103 (8%) and 8 for O5 (7%)

serogroups. The analysis of the draft sequences of 15 O5 AE-STEC could not

identify any difference correlated to the host. As a conclusion, (i) the AE-STEC

associated with diarrhoea in young calves still belong to the same serogroups as

previously (O5, O26, O111) but the O103 serogroup may be emerging, (ii) the O5

AE-STEC from calves and humans are genetically similar.

Copyright © 2016. Published by Elsevier B.V.

DOI: 10.1016/j.vetmic.2016.02.017

PMID: 26923249

332. J Chem Inf Model. 2016 Feb 22;56(2):423-34. doi: 10.1021/acs.jcim.5b00517. Epub

2016 Feb 5.

Accurate Prediction of Contact Numbers for Multi-Spanning Helical Membrane

Proteins.

Li B(1,)(2), Mendenhall J(1,)(2), Nguyen ED(2), Weiner BE(1,)(2), Fischer

AW(1,)(2), Meiler J(1,)(2).

Author information:

(1)Department of Chemistry, Vanderbilt University , Nashville, Tennessee 37232,

United States. (2)Center for Structural Biology, Vanderbilt University ,

Nashville, Tennessee 37232, United States.

Prediction of the three-dimensional (3D) structures of proteins by computational

methods is acknowledged as an unsolved problem. Accurate prediction of important

structural characteristics such as contact number is expected to accelerate the

otherwise slow progress being made in the prediction of 3D structure of proteins.

Here, we present a dropout neural network-based method, TMH-Expo, for predicting

the contact number of transmembrane helix (TMH) residues from sequence. Neuronal

dropout is a strategy where certain neurons of the network are excluded from

back-propagation to prevent co-adaptation of hidden-layer neurons. By using

neuronal dropout, overfitting was significantly reduced and performance was

noticeably improved. For multi-spanning helical membrane proteins, TMH-Expo

achieved a remarkable Pearson correlation coefficient of 0.69 between predicted

and experimental values and a mean absolute error of only 1.68. In addition,

among those membrane protein-membrane protein interface residues, 76.8% were

correctly predicted. Mapping of predicted contact numbers onto structures

indicates that contact numbers predicted by TMH-Expo reflect the exposure

patterns of TMHs and reveal membrane protein-membrane protein interfaces,

reinforcing the potential of predicted contact numbers to be used as restraints

for 3D structure prediction and protein-protein docking. TMH-Expo can be accessed

via a Web server at www.meilerlab.org .

DOI: 10.1021/acs.jcim.5b00517

PMID: 26804342 [Indexed for MEDLINE]

333. J Mol Biol. 2016 Feb 22;428(4):709-19. doi: 10.1016/j.jmb.2016.01.029. Epub 2016

Feb 5.

CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction

of Cryptic Binding Sites.

Cimermancic P(1), Weinkam P(2), Rettenmaier TJ(3), Bichmann L(2), Keedy DA(2),

Woldeyes RA(4), Schneidman-Duhovny D(2), Demerdash ON(5), Mitchell JC(6), Wells

JA(7), Fraser JS(2), Sali A(8).

Author information:

(1)Department of Bioengineering and Therapeutic Sciences, University of

California, San Francisco, San Francisco, CA 94158, USA; Graduate Group in

Biological and Medical Informatics,University of California, San Francisco, San

Francisco, CA 94158, USA. Electronic address: peter.cimermancic@ucsf.edu.

(2)Department of Bioengineering and Therapeutic Sciences, University of

California, San Francisco, San Francisco, CA 94158, USA. (3)Graduate Group in

Chemistry and Chemical Biology, University of California, San Francisco, San

Francisco, CA 94158, USA; Pharmaceutical Chemistry and California Institute for

Quantitative Biosciences, University of California, San Francisco, San Francisco,

CA 94158, USA. (4)Department of Bioengineering and Therapeutic Sciences,

University of California, San Francisco, San Francisco, CA 94158, USA; Graduate

Group in Chemistry and Chemical Biology, University of California, San Francisco,

San Francisco, CA 94158, USA. (5)Department of Chemical and Biomolecular

Engineering, University of California, Berkeley, Berkeley, CA 94720, USA.

(6)Departments of Biochemistry and Mathematics, University of Wisconsin-Madison,

Madison, WI 53706, USA. (7)Pharmaceutical Chemistry and California Institute for

Quantitative Biosciences, University of California, San Francisco, San Francisco,

CA 94158, USA; Cellular and Molecular Pharmacology and California Institute for

Quantitative Biosciences, University of California, San Francisco, San Francisco,

CA 94158, USA. (8)Department of Bioengineering and Therapeutic Sciences,

University of California, San Francisco, San Francisco, CA 94158, USA;

Pharmaceutical Chemistry and California Institute for Quantitative Biosciences,

University of California, San Francisco, San Francisco, CA 94158, USA. Electronic

address: http://salilab.org.

Many proteins have small-molecule binding pockets that are not easily detectable

in the ligand-free structures. These cryptic sites require a conformational

change to become apparent; a cryptic site can therefore be defined as a site that

forms a pocket in a holo structure, but not in the apo structure. Because many

proteins appear to lack druggable pockets, understanding and accurately

identifying cryptic sites could expand the set of drug targets. Previously,

cryptic sites were identified experimentally by fragment-based ligand discovery

and computationally by long molecular dynamics simulations and fragment docking.

Here, we begin by constructing a set of structurally defined apo-holo pairs with

cryptic sites. Next, we comprehensively characterize the cryptic sites in terms

of their sequence, structure, and dynamics attributes. We find that cryptic sites

tend to be as conserved in evolution as traditional binding pockets but are less

hydrophobic and more flexible. Relying on this characterization, we use machine

learning to predict cryptic sites with relatively high accuracy (for our

benchmark, the true positive and false positive rates are 73% and 29%,

respectively). We then predict cryptic sites in the entire structurally

characterized human proteome (11,201 structures, covering 23% of all residues in

the proteome). CryptoSite increases the size of the potentially "druggable" human

proteome from ~40% to ~78% of disease-associated proteins. Finally, to

demonstrate the utility of our approach in practice, we experimentally validate a

cryptic site in protein tyrosine phosphatase 1B using a covalent ligand and NMR

spectroscopy. The CryptoSite Web server is available at

http://salilab.org/cryptosite.

Published by Elsevier Ltd.

DOI: 10.1016/j.jmb.2016.01.029

PMCID: PMC4794384

PMID: 26854760 [Indexed for MEDLINE]

334. J Mol Biol. 2016 Feb 22;428(4):702-8. doi: 10.1016/j.jmb.2015.10.017. Epub 2015

Oct 27.

PhyreStorm: A Web Server for Fast Structural Searches Against the PDB.

Mezulis S(1), Sternberg MJ(2), Kelley LA(2).

Author information:

(1)Structural Bioinformatics Group, Imperial College London, London SW7 2AZ,

United Kingdom. Electronic address: stefans.mezulis08@imperial.ac.uk.

(2)Structural Bioinformatics Group, Imperial College London, London SW7 2AZ,

United Kingdom.

The identification of structurally similar proteins can provide a range of

biological insights, and accordingly, the alignment of a query protein to a

database of experimentally determined protein structures is a technique commonly

used in the fields of structural and evolutionary biology. The PhyreStorm Web

server has been designed to provide comprehensive, up-to-date and rapid

structural comparisons against the Protein Data Bank (PDB) combined with a rich

and intuitive user interface. It is intended that this facility will enable

biologists inexpert in bioinformatics access to a powerful tool for exploring

protein structure relationships beyond what can be achieved by sequence analysis

alone. By partitioning the PDB into similar structures, PhyreStorm is able to

quickly discard the majority of structures that cannot possibly align well to a

query protein, reducing the number of alignments required by an order of

magnitude. PhyreStorm is capable of finding 93±2% of all highly similar

(TM-score>0.7) structures in the PDB for each query structure, usually in less

than 60s. PhyreStorm is available at http://www.sbg.bio.ic.ac.uk/phyrestorm/.

Copyright © 2015 The Authors. Published by Elsevier Ltd.. All rights reserved.

DOI: 10.1016/j.jmb.2015.10.017

PMID: 26517951 [Indexed for MEDLINE]

335. J Mol Biol. 2016 Feb 22;428(4):720-5. doi: 10.1016/j.jmb.2015.09.014. Epub 2015

Sep 26.

The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular

Complexes.

van Zundert GC(1), Rodrigues JP(1), Trellet M(2), Schmitz C(3), Kastritis PL(4),

Karaca E(4), Melquiond AS(5), van Dijk M(6), de Vries SJ(7), Bonvin AM(1).

Author information:

(1)Bijvoet Center for Biomolecular Research, Faculty of Science Department of

Chemistry, Utrecht University, Domplein 29, 3512 JE Utrecht, the Netherlands.

(2)Centre National de la Recherche Scientifique Laboratoire d'Informatique pour

la Mécanique et les Sciences de l'Ingénieur, rue John Von Neumann, 91403 Orsay,

France. (3)Instaclustr Level 5, 1 Moore Street, Canberra ACT 2600, Australia.

(4)European Molecular Biology Laboratory Heidelberg, Meyerhofstrasse 1, 69117

Heidelberg, Germany. (5)Hubrecht Institute, Uppsalalaan 8, 3584 CT Utrecht, the

Netherlands. (6)Faculty of Sciences, VU University Amsterdam, De Boelelaan 1083,

1081 HV Amsterdam, the Netherlands. (7)Physik-Department, Technische Universität

München, James-Franck-Strasse 1, 85748 Garching, Germany.

The prediction of the quaternary structure of biomolecular macromolecules is of

paramount importance for fundamental understanding of cellular processes and drug

design. In the era of integrative structural biology, one way of increasing the

accuracy of modeling methods used to predict the structure of biomolecular

complexes is to include as much experimental or predictive information as

possible in the process. This has been at the core of our information-driven

docking approach HADDOCK. We present here the updated version 2.2 of the HADDOCK

portal, which offers new features such as support for mixed molecule types,

additional experimental restraints and improved protocols, all of this in a

user-friendly interface. With well over 6000 registered users and 108,000 jobs

served, an increasing fraction of which on grid resources, we hope that this

timely upgrade will help the community to solve important biological questions

and further advance the field. The HADDOCK2.2 Web server is freely accessible to

non-profit users at http://haddock.science.uu.nl/services/HADDOCK2.2.

Copyright © 2015 The Authors. Published by Elsevier Ltd.. All rights reserved.

DOI: 10.1016/j.jmb.2015.09.014

PMID: 26410586 [Indexed for MEDLINE]

336. Database (Oxford). 2016 Feb 17;2016. pii: bav119. doi: 10.1093/database/bav119.

Print 2016.

Ensembl regulation resources.

Zerbino DR(1), Johnson N(2), Juetteman T(2), Sheppard D(3), Wilder SP(2), Lavidas

I(2), Nuhn M(2), Perry E(2), Raffaillac-Desfosses Q(2), Sobral D(2), Keefe D(2),

Gräf S(2), Ahmed I(2), Kinsella R(2), Pritchard B(4), Brent S(4), Amode R(3),

Parker A(3), Trevanion S(3), Birney E(2), Dunham I(2), Flicek P(3).

Author information:

(1)European Molecular Biology Laboratory, European Bioinformatics Institute,

Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK zerbino@ebi.ac.uk.

(2)European Molecular Biology Laboratory, European Bioinformatics Institute,

Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. (3)European

Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust

Genome Campus, Hinxton, Cambridge CB10 1SD, UK Wellcome Trust Sanger Institute,

Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. (4)Wellcome Trust

Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

New experimental techniques in epigenomics allow researchers to assay a diversity

of highly dynamic features such as histone marks, DNA modifications or chromatin

structure. The study of their fluctuations should provide insights into gene

expression regulation, cell differentiation and disease. The Ensembl project

collects and maintains the Ensembl regulation data resources on epigenetic marks,

transcription factor binding and DNA methylation for human and mouse, as well as

microarray probe mappings and annotations for a variety of chordate genomes. From

this data, we produce a functional annotation of the regulatory elements along

the human and mouse genomes with plans to expand to other species as data becomes

available. Starting from well-studied cell lines, we will progressively expand

our library of measurements to a greater variety of samples. Ensembl's regulation

resources provide a central and easy-to-query repository for reference

epigenomes. As with all Ensembl data, it is freely available at

http://www.ensembl.org, from the Perl and REST APIs and from the public Ensembl

MySQL database server at ensembldb.ensembl.org. Database URL:

http://www.ensembl.org.

© The Author(s) 2016. Published by Oxford University Press.

DOI: 10.1093/database/bav119

PMCID: PMC4756621

PMID: 26888907 [Indexed for MEDLINE]

337. Sci Rep. 2016 Feb 17;6:21280. doi: 10.1038/srep21280.

HIV coreceptor tropism determination and mutational pattern identification.

Shen HS(1), Yin J(2), Leng F(1), Teng RF(1), Xu C(1), Xia XY(1), Pan XM(1).

Author information:

(1)The Key Laboratory of Bioinformatics, Ministry of Education, School of Life

Sciences, Tsinghua University, China. (2)Department of Biostatistics, Saw Swee

Hock School of Public Health, National University of Singapore, Singapore.

In the early stages of infection, Human Immunodeficiency Virus Type 1 (HIV-1)

generally selects CCR5 as the primary coreceptor for entering the host cell. As

infection progresses, the virus evolves and may exhibit a coreceptor-switch to

CXCR4. Accurate determination coreceptor usage and identification key mutational

patterns associated tropism switch are essential for selection of appropriate

therapies and understanding mechanism of coreceptor change. We developed a

classifier composed of two coreceptor-specific weight matrices (CMs) based on a

full-scale dataset. For this classifier, we found an AUC of 0.97, an accuracy of

95.21% and an MCC of 0.885 (sensitivity 92.92%; specificity 95.54%) in a ten-fold

cross-validation, outperforming all other methods on an independent dataset (13%

higher MCC value than geno2pheno and 15% higher MCC value than PSSM). A web

server (http://spg.med.tsinghua.edu.cn/CM.html) based on our classifier was

provided. Patterns of genetic mutations that occur along with coreceptor

transitions were further identified based on the score of each sequence. Six

pairs of one-AA mutational patterns and three pairs of two-AA mutational patterns

were identified to associate with increasing propensity for X4 tropism. These

mutational patterns offered new insights into the mechanism of coreceptor switch

and aided in monitoring coreceptor switch.

DOI: 10.1038/srep21280

PMCID: PMC4756667

PMID: 26883082 [Indexed for MEDLINE]

338. Bioinformatics. 2016 Feb 15;32(4):614-5. doi: 10.1093/bioinformatics/btv607. Epub

2015 Oct 29.

DelPhiPKa web server: predicting pKa of proteins, RNAs and DNAs.

Wang L(1), Zhang M(2), Alexov E(1).

Author information:

(1)Computational Biophysics and Bioinformatics, Department of Physics and

Astronomy, Clemson University, Clemson, SC 29634, USA and. (2)Department of

Computer Sciences, East Carolina University, Greenville, NC 27858, USA.

A new pKa prediction web server is released, which implements DelPhi Gaussian

dielectric function to calculate electrostatic potentials generated by charges of

biomolecules. Topology parameters are extended to include atomic information of

nucleotides of RNA and DNA, which extends the capability of pKa calculations

beyond proteins. The web server allows the end-user to protonate the biomolecule

at particular pH based on calculated pKa values and provides the downloadable

file in PQR format. Several tests are performed to benchmark the accuracy and

speed of the protocol.IMPLEMENTATION: The web server follows a client-server

architecture built on PHP and HTML and utilizes DelPhiPKa program. The

computation is performed on the Palmetto supercomputer cluster and

results/download links are given back to the end-user via http protocol. The web

server takes advantage of MPI parallel implementation in DelPhiPKa and can run a

single job on up to 24 CPUs.

AVAILABILITY AND IMPLEMENTATION: The DelPhiPKa web server is available at

http://compbio.clemson.edu/pka\_webserver.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv607

PMID: 26515825 [Indexed for MEDLINE]

339. Bioinformatics. 2016 Feb 15;32(4):619-20. doi: 10.1093/bioinformatics/btv614.

Epub 2015 Oct 27.

Omokage search: shape similarity search service for biomolecular structures in

both the PDB and EMDB.

Suzuki H(1), Kawabata T(1), Nakamura H(1).

Author information:

(1)Institute for Protein Research, Osaka University, Suita, Osaka, 565-0871,

Japan.

Omokage search is a service to search the global shape similarity of biological

macromolecules and their assemblies, in both the Protein Data Bank (PDB) and

Electron Microscopy Data Bank (EMDB). The server compares global shapes of

assemblies independent of sequence order and number of subunits. As a search

query, the user inputs a structure ID (PDB ID or EMDB ID) or uploads an atomic

model or 3D density map to the server. The search is performed usually within

1 min, using one-dimensional profiles (incremental distance rank profiles) to

characterize the shapes. Using the gmfit (Gaussian mixture model fitting)

program, the found structures are fitted onto the query structure and their

superimposed structures are displayed on the Web browser. Our service provides

new structural perspectives to life science researchers.AVAILABILITY AND

IMPLEMENTATION: Omokage search is freely accessible at http://pdbj.org/omokage/.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv614

PMCID: PMC4743628

PMID: 26508754 [Indexed for MEDLINE]

340. Bioinformatics. 2016 Feb 15;32(4):616-8. doi: 10.1093/bioinformatics/btv611. Epub

2015 Oct 25.

DARA: a web server for rapid search of structural neighbours using solution small

angle X-ray scattering data.

Kikhney AG(1), Panjkovich A(1), Sokolova AV(2), Svergun DI(1).

Author information:

(1)European Molecular Biology Laboratory, Hamburg Outstation, EMBL c/o DESY,

22607 Hamburg, Germany and. (2)Bragg Institute, ANSTO, Lucas Heights, 2234 NSW

Australia.

MOTIVATION: Small angle X-ray scattering (SAXS) is an established method for

studying biological macromolecules in solution, whereby the experimental

scattering patterns relate to the quaternary and tertiary structure of the

macromolecule. Here we present DARA, a web-server, that queries over 150 000

scattering profiles pre-computed from the high resolution models of

macromolecules and biological assemblies in the Protein Data Bank, to rapidly

find nearest neighbours of a given experimental or theoretical SAXS pattern.

Identification of the best scattering equivalents provides a straightforward and

automated way of structural assessment of macromolecules based on a SAXS profile.

DARA results are useful e.g. for fold recognition and finding of biologically

active oligomers.

AVAILABILITY AND IMPLEMENTATION: http://dara.embl-hamburg.de/.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv611

PMCID: PMC4743626

PMID: 26504146 [Indexed for MEDLINE]

341. Bioinformatics. 2016 Feb 15;32(4):611-3. doi: 10.1093/bioinformatics/btv595. Epub

2015 Oct 26.

NMRe: a web server for NMR protein structure refinement with high-quality

structure validation scores.

Ryu H(1), Lim G(1), Sung BH(2), Lee J(1).

Author information:

(1)Korean Bioinformation Center (KOBIC), Korea Research Institute of Bioscience

and Biotechnology, Yuseong-gu, Daejeon 305-806, the Republic of Korea, Department

of Nanobiotechnology and Bioinformatics, University of Science and Technology,

Yuseong-Gu, Daejeon 305-350, the Republic of Korea and. (2)Bioenergy and

Biochemical Research Center, Korea Research Institute of Bioscience and

Biotechnology, Yuseong-gu, Daejeon 305-806, the Republic of Korea.

Protein structure refinement is a necessary step for the study of protein

function. In particular, some nuclear magnetic resonance (NMR) structures are of

lower quality than X-ray crystallographic structures. Here, we present NMRe, a

web-based server for NMR structure refinement. The previously developed

knowledge-based energy function STAP (Statistical Torsion Angle Potential) was

used for NMRe refinement. With STAP, NMRe provides two refinement protocols using

two types of distance restraints. If a user provides NOE (Nuclear Overhauser

Effect) data, the refinement is performed with the NOE distance restraints as a

conventional NMR structure refinement. Additionally, NMRe generates NOE-like

distance restraints based on the inter-hydrogen distances derived from the input

structure. The efficiency of NMRe refinement was validated on 20 NMR structures.

Most of the quality assessment scores of the refined NMR structures were better

than those of the original structures. The refinement results are provided as a

three-dimensional structure view, a secondary structure scheme, and numerical and

graphical structure validation scores.AVAILABILITY AND IMPLEMENTATION: NMRe is

available at http://psb.kobic.re.kr/nmre/.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv595

PMID: 26504145 [Indexed for MEDLINE]

342. PLoS One. 2016 Feb 12;11(2):e0149350. doi: 10.1371/journal.pone.0149350.

eCollection 2016.

RV-Typer: A Web Server for Typing of Rhinoviruses Using Alignment-Free Approach.

Kolekar PS(1), Waman VP(1), Kale MM(2), Kulkarni-Kale U(1).

Author information:

(1)Bioinformatics Centre, Savitribai Phule Pune University (formerly University

of Pune), Pune, 411 007, India. (2)Department of Statistics, Savitribai Phule

Pune University (formerly University of Pune), Pune, 411 007, India.

Rhinoviruses (RV) are increasingly being reported to cause mild to severe

infections of respiratory tract in humans. RV are antigenically the most diverse

species of the genus Enterovirus and family Picornaviridae. There are three

species of RV (RV-A, -B and -C), with 80, 32 and 55 serotypes/types,

respectively. Antigenic variation is the main limiting factor for development of

a cross-protective vaccine against RV.Serotyping of Rhinoviruses is carried out

using cross-neutralization assays in cell culture. However, these assays become

laborious and time-consuming for the large number of strains. Alternatively,

serotyping of RV is carried out by alignment-based phylogeny of both protein and

nucleotide sequences of VP1. However, serotyping of RV based on alignment-based

phylogeny is a multi-step process, which needs to be repeated every time a new

isolate is sequenced. In view of the growing need for serotyping of RV, an

alignment-free method based on "return time distribution" (RTD) of amino acid

residues in VP1 protein has been developed and implemented in the form of a web

server titled RV-Typer. RV-Typer accepts nucleotide or protein sequences as an

input and computes return times of di-peptides (k = 2) to assign serotypes. The

RV-Typer performs with 100% sensitivity and specificity. It is significantly

faster than alignment-based methods. The web server is available at

http://bioinfo.net.in/RV-Typer/home.html.

DOI: 10.1371/journal.pone.0149350

PMCID: PMC4752186

PMID: 26870949 [Indexed for MEDLINE]

343. Sci Rep. 2016 Feb 10;6:20678. doi: 10.1038/srep20678.

Prediction of Immunomodulatory potential of an RNA sequence for designing

non-toxic siRNAs and RNA-based vaccine adjuvants.

Chaudhary K(1), Nagpal G(1), Dhanda SK(1), Raghava GP(1).

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India.

Our innate immune system recognizes a foreign RNA sequence of a pathogen and

activates the immune system to eliminate the pathogen from our body. This

immunomodulatory potential of RNA can be used to design RNA-based immunotherapy

and vaccine adjuvants. In case of siRNA-based therapy, the immunomodulatory

effect of an RNA sequence is unwanted as it may cause immunotoxicity. Thus, we

developed a method for designing a single-stranded RNA (ssRNA) sequence with

desired immunomodulatory potentials, for designing RNA-based therapeutics,

immunotherapy and vaccine adjuvants. The dataset used for training and testing

our models consists of 602 experimentally verified immunomodulatory

oligoribonucleotides (IMORNs) that are ssRNA sequences of length 17 to 27

nucleotides and 520 circulating miRNAs as non-immunomodulatory sequences. We

developed prediction models using various features that include composition-based

features, binary profile, selected features, and hybrid features. All models were

evaluated using five-fold cross-validation and external validation techniques;

achieving a maximum mean Matthews Correlation Coefficient (MCC) of 0.86 with 93%

accuracy. We identified motifs using MERCI software and observed the abundance of

adenine (A) in motifs. Based on the above study, we developed a web server,

imRNA, comprising of various modules important for designing RNA-based

therapeutics (http://crdd.osdd.net/raghava/imrna/).

DOI: 10.1038/srep20678

PMCID: PMC4748260

PMID: 26861761 [Indexed for MEDLINE]

344. BMC Cancer. 2016 Feb 9;16:77. doi: 10.1186/s12885-016-2082-y.

Prediction of anticancer molecules using hybrid model developed on molecules

screened against NCI-60 cancer cell lines.

Singh H(1), Kumar R(2), Singh S(3), Chaudhary K(4), Gautam A(5), Raghava GP(6).

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A,

Chandigarh, India. harinder@imtech.res.in. (2)Bioinformatics Centre, Institute of

Microbial Technology, Sector 39-A, Chandigarh, India. rahulk@imtech.res.in.

(3)Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A,

Chandigarh, India. sandysingh@imtech.res.in. (4)Bioinformatics Centre, Institute

of Microbial Technology, Sector 39-A, Chandigarh, India. kumardeep@imtech.res.in.

(5)Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A,

Chandigarh, India. ankurgautam@imtech.res.in. (6)Bioinformatics Centre, Institute

of Microbial Technology, Sector 39-A, Chandigarh, India. raghava@imtech.res.in.

BACKGROUND: In past, numerous quantitative structure-activity relationship (QSAR)

based models have been developed for predicting anticancer activity for a

specific class of molecules against different cancer drug targets. In contrast,

limited attempt have been made to predict the anticancer activity of a diverse

class of chemicals against a wide variety of cancer cell lines. In this study, we

described a hybrid method developed on thousands of anticancer and non-anticancer

molecules tested against National Cancer Institute (NCI) 60 cancer cell lines.

RESULTS: Our analysis of anticancer molecules revealed that majority of

anticancer molecules contains 18-24 carbon atoms and are dominated by functional

groups like R2NH, R3N, ROH, RCOR, and ROR. It was also observed that certain

substructures (e.g., 1-methoxy-4-methylbenzene, 1-methoxy benzene, Nitrobenzene,

Indole, Propenyl benzene) are more abundant in anticancer molecules. Next, we

developed anticancer molecule prediction models using various machine-learning

techniques and achieved maximum matthews correlation coefficient (MCC) of 0.81

with 90.40% accuracy using support vector machine (SVM) based models. In another

approach, a novel similarity or potency score based method has been developed

using selected fragments/fingerprints and achieved maximum MCC of 0.82 with

90.65% accuracy. Finally, we combined the strength of above methods and developed

a hybrid method with maximum MCC of 0.85 with 92.47% accuracy.

CONCLUSIONS: We developed a hybrid method utilizing the best of machine learning

and potency score based method. The highly accurate hybrid method can be used for

classification of anticancer and non-anticancer molecules. In order to facilitate

scientific community working in the field of anticancer drug discovery, we

integrate hybrid and potency method in a web server CancerIN. This server

provides various facilities that includes; virtual screening of anticancer

molecules, analog based drug design, and similarity with known anticancer

molecules ( http://crdd.osdd.net/oscadd/cancerin).

DOI: 10.1186/s12885-016-2082-y

PMCID: PMC4748564

PMID: 26860193 [Indexed for MEDLINE]

345. PLoS One. 2016 Feb 9;11(2):e0148321. doi: 10.1371/journal.pone.0148321.

eCollection 2016.

ePIANNO: ePIgenomics ANNOtation tool.

Liu CH(1,)(2,)(3), Ho BC(4,)(5), Chen CL(1), Chang YH(1), Hsu YC(1), Li YC(1),

Yuan SS(1), Huang YH(1), Chang CS(5), Li KC(1), Chen HY(1).

Author information:

(1)Institute of Statistical Science, Academia Sinica, Nangang, Taipei, Taiwan.

(2)Bioinformatics Program, Taiwan International Graduate Program, Academia

Sinica, Nangang, Taipei, Taiwan. (3)Institute of Biomedical Informatics, National

Yang-Ming University, Taipei, Taiwan. (4)Department of Clinical Laboratory

Sciences and Medical Biotechnology, College of Medicine, National Taiwan

University, Taipei, Taiwan. (5)NTU Center for Genomic Medicine, National Taiwan

University College of Medicine, Taipei, Taiwan.

Recently, with the development of next generation sequencing (NGS), the

combination of chromatin immunoprecipitation (ChIP) and NGS, namely ChIP-seq, has

become a powerful technique to capture potential genomic binding sites of

regulatory factors, histone modifications and chromatin accessible regions. For

most researchers, additional information including genomic variations on the TF

binding site, allele frequency of variation between different populations,

variation associated disease, and other neighbour TF binding sites are essential

to generate a proper hypothesis or a meaningful conclusion. Many ChIP-seq

datasets had been deposited on the public domain to help researchers make new

discoveries. However, researches are often intimidated by the complexity of data

structure and largeness of data volume. Such information would be more useful if

they could be combined or downloaded with ChIP-seq data. To meet such demands, we

built a webtool: ePIgenomic ANNOtation tool (ePIANNO,

http://epianno.stat.sinica.edu.tw/index.html). ePIANNO is a web server that

combines SNP information of populations (1000 Genomes Project) and gene-disease

association information of GWAS (NHGRI) with ChIP-seq (hmChIP, ENCODE, and

ROADMAP epigenomics) data. ePIANNO has a user-friendly website interface allowing

researchers to explore, navigate, and extract data quickly. We use two examples

to demonstrate how users could use functions of ePIANNO webserver to explore

useful information about TF related genomic variants. Users could use our query

functions to search target regions, transcription factors, or annotations.

ePIANNO may help users to generate hypothesis or explore potential biological

functions for their studies.

DOI: 10.1371/journal.pone.0148321

PMCID: PMC4747527

PMID: 26859295 [Indexed for MEDLINE]

346. Acta Crystallogr D Struct Biol. 2016 Feb;72(Pt 2):266-80. doi:

10.1107/S2059798315024730. Epub 2016 Jan 28.

Fitmunk: improving protein structures by accurate, automatic modeling of

side-chain conformations.

Porebski PJ(1), Cymborowski M(1), Pasenkiewicz-Gierula M(2), Minor W(1).

Author information:

(1)Department of Molecular Physiology and Biological Physics, University of

Virginia, Jordan Hall, 1340 Jefferson Park Avenue, Charlottesville, VA 22908,

USA. (2)Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian

University, ul. Gronostajowa 7, 30-387 Kraków, Poland.

Improvements in crystallographic hardware and software have allowed automated

structure-solution pipelines to approach a near-`one-click' experience for the

initial determination of macromolecular structures. However, in many cases the

resulting initial model requires a laborious, iterative process of refinement and

validation. A new method has been developed for the automatic modeling of

side-chain conformations that takes advantage of rotamer-prediction methods in a

crystallographic context. The algorithm, which is based on deterministic dead-end

elimination (DEE) theory, uses new dense conformer libraries and a hybrid energy

function derived from experimental data and prior information about rotamer

frequencies to find the optimal conformation of each side chain. In contrast to

existing methods, which incorporate the electron-density term into

protein-modeling frameworks, the proposed algorithm is designed to take advantage

of the highly discriminatory nature of electron-density maps. This method has

been implemented in the program Fitmunk, which uses extensive conformational

sampling. This improves the accuracy of the modeling and makes it a versatile

tool for crystallographic model building, refinement and validation. Fitmunk was

extensively tested on over 115 new structures, as well as a subset of 1100

structures from the PDB. It is demonstrated that the ability of Fitmunk to model

more than 95% of side chains accurately is beneficial for improving the quality

of crystallographic protein models, especially at medium and low resolutions.

Fitmunk can be used for model validation of existing structures and as a tool to

assess whether side chains are modeled optimally or could be better fitted into

electron density. Fitmunk is available as a web service at

http://kniahini.med.virginia.edu/fitmunk/server/ or at

http://fitmunk.bitbucket.org/.

DOI: 10.1107/S2059798315024730

PMCID: PMC4756610

PMID: 26894674 [Indexed for MEDLINE]

347. Bioinformatics. 2016 Feb 1;32(3):362-9. doi: 10.1093/bioinformatics/btv604. Epub

2015 Oct 17.

iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength

by pseudo k-tuple nucleotide composition.

Liu B(1), Fang L(2), Long R(2), Lan X(3), Chou KC(4).

Author information:

(1)School of Computer Science and Technology, Key Laboratory of Network Oriented

Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School,

Shenzhen, Guangdong 518055, China, Computational Biology, Gordon Life Science

Institute, Belmont, MA 02478, USA. (2)School of Computer Science and Technology.

(3)Department of Genetics, Stanford University, Stanford, CA 94305, USA and.

(4)Computational Biology, Gordon Life Science Institute, Belmont, MA 02478, USA,

Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz

University, Jeddah 21589, Saudi Arabia.

MOTIVATION: Enhancers are of short regulatory DNA elements. They can be bound

with proteins (activators) to activate transcription of a gene, and hence play a

critical role in promoting gene transcription in eukaryotes. With the avalanche

of DNA sequences generated in the post-genomic age, it is a challenging task to

develop computational methods for timely identifying enhancers from extremely

complicated DNA sequences. Although some efforts have been made in this regard,

they were limited at only identifying whether a query DNA element being of an

enhancer or not. According to the distinct levels of biological activities and

regulatory effects on target genes, however, enhancers should be further

classified into strong and weak ones in strength.

RESULTS: In view of this, a two-layer predictor called ' IENHANCER-2L: ' was

proposed by formulating DNA elements with the 'pseudo k-tuple nucleotide

composition', into which the six DNA local parameters were incorporated. To the

best of our knowledge, it is the first computational predictor ever established

for identifying not only enhancers, but also their strength. Rigorous

cross-validation tests have indicated that IENHANCER-2L: holds very high

potential to become a useful tool for genome analysis.

AVAILABILITY AND IMPLEMENTATION: For the convenience of most experimental

scientists, a web server for the two-layer predictor was established at

http://bioinformatics.hitsz.edu.cn/iEnhancer-2L/, by which users can easily get

their desired results without the need to go through the mathematical details.

CONTACT: bliu@gordonlifescience.org, bliu@insun.hit.edu.cn, xlan@stanford.edu,

kcchou@gordonlifescience.org

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv604

PMID: 26476782 [Indexed for MEDLINE]

348. Bioinformatics. 2016 Feb 1;32(3):370-7. doi: 10.1093/bioinformatics/btv580. Epub

2015 Oct 10.

Fast and accurate non-sequential protein structure alignment using a new

asymmetric linear sum assignment heuristic.

Brown P(1), Pullan W(1), Yang Y(2), Zhou Y(3).

Author information:

(1)School of ICT, Griffith University, Gold Coast, QLD 4222, Australia.

(2)Institute for Glycomics, Griffith University, Gold Coast, QLD 4222, Australia.

(3)School of ICT, Griffith University, Gold Coast, QLD 4222, Australia Institute

for Glycomics, Griffith University, Gold Coast, QLD 4222, Australia.

MOTIVATION: The three dimensional tertiary structure of a protein at near atomic

level resolution provides insight alluding to its function and evolution. As

protein structure decides its functionality, similarity in structure usually

implies similarity in function. As such, structure alignment techniques are often

useful in the classifications of protein function. Given the rapidly growing rate

of new, experimentally determined structures being made available from

repositories such as the Protein Data Bank, fast and accurate computational

structure comparison tools are required. This paper presents SPalignNS, a

non-sequential protein structure alignment tool using a novel asymmetrical greedy

search technique.

RESULTS: The performance of SPalignNS was evaluated against existing sequential

and non-sequential structure alignment methods by performing trials with commonly

used datasets. These benchmark datasets used to gauge alignment accuracy include

(i) 9538 pairwise alignments implied by the HOMSTRAD database of homologous

proteins; (ii) a subset of 64 difficult alignments from set (i) that have low

structure similarity; (iii) 199 pairwise alignments of proteins with similar

structure but different topology; and (iv) a subset of 20 pairwise alignments

from the RIPC set. SPalignNS is shown to achieve greater alignment accuracy

(lower or comparable root-mean squared distance with increased structure overlap

coverage) for all datasets, and the highest agreement with reference alignments

from the challenging dataset (iv) above, when compared with both sequentially

constrained alignments and other non-sequential alignments.

AVAILABILITY AND IMPLEMENTATION: SPalignNS was implemented in C++. The source

code, binary executable, and a web server version is freely available at:

http://sparks-lab.org

CONTACT: yaoqi.zhou@griffith.edu.au.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv580

PMID: 26454279 [Indexed for MEDLINE]

349. Bioinformatics. 2016 Feb 1;32(3):462-4. doi: 10.1093/bioinformatics/btv581. Epub

2015 Oct 10.

FALCON@home: a high-throughput protein structure prediction server based on

remote homologue recognition.

Wang C(1), Zhang H(1), Zheng WM(2), Xu D(3), Zhu J(4), Wang B(4), Ning K(5), Sun

S(4), Li SC(6), Bu D(4).

Author information:

(1)Key Lab of Intelligent Information Processing, Institute of Computing

Technology, Chinese Academy of Sciences, Beijing, China, University of Chinese

Academy of Sciences, Beijing, China. (2)Institute of Theoretical Physics, Chinese

Academy of Sciences, Beijing, China. (3)Department of Computer Science and

Christopher S. Bond Life Sciences Center, University of Missouri, MO 65211, USA.

(4)Key Lab of Intelligent Information Processing, Institute of Computing

Technology, Chinese Academy of Sciences, Beijing, China. (5)College of Life

Science, Huazhong University of Science and Technology, Wuhan, China and.

(6)Department of Computer Science, City University of Hong Kong, Kowloon, Hong

Kong.

SUMMARY: The protein structure prediction approaches can be categorized into

template-based modeling (including homology modeling and threading) and free

modeling. However, the existing threading tools perform poorly on remote

homologous proteins. Thus, improving fold recognition for remote homologous

proteins remains a challenge. Besides, the proteome-wide structure prediction

poses another challenge of increasing prediction throughput. In this study, we

presented FALCON@home as a protein structure prediction server focusing on remote

homologue identification. The design of FALCON@home is based on the observation

that a structural template, especially for remote homologous proteins, consists

of conserved regions interweaved with highly variable regions. The highly

variable regions lead to vague alignments in threading approaches. Thus,

FALCON@home first extracts conserved regions from each template and then aligns a

query protein with conserved regions only rather than the full-length template

directly. This helps avoid the vague alignments rooted in highly variable

regions, improving remote homologue identification. We implemented FALCON@home

using the Berkeley Open Infrastructure of Network Computing (BOINC) volunteer

computing protocol. With computation power donated from over 20,000 volunteer

CPUs, FALCON@home shows a throughput as high as processing of over 1000 proteins

per day. In the Critical Assessment of protein Structure Prediction (CASP11), the

FALCON@home-based prediction was ranked the 12th in the template-based modeling

category. As an application, the structures of 880 mouse mitochondria proteins

were predicted, which revealed the significant correlation between protein

half-lives and protein structural factors.

AVAILABILITY AND IMPLEMENTATION: FALCON@home is freely available at

http://protein.ict.ac.cn/FALCON/.

CONTACT: shuaicli@cityu.edu.hk, dbu@ict.ac.cn

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv581

PMCID: PMC4804767

PMID: 26454278 [Indexed for MEDLINE]

350. Bioinformatics. 2016 Feb 1;32(3):474-6. doi: 10.1093/bioinformatics/btv574. Epub

2015 Oct 6.

The new protein topology graph library web server.

Schäfer T(1), Scheck A(1), Bruneß D(1), May P(2), Koch I(1).

Author information:

(1)Molecular Bioinformatics, Institute of Computer Science, Johann Wolfgang

Goethe-University Frankfurt am Main, Robert-Mayer-Str. 11-15, 60325 Frankfurt am

Main, Germany and. (2)Genome Analysis, Luxembourg Centre for Systems Biomedicine,

University of Luxembourg, 7, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette,

Luxembourg.

SUMMARY: We present a new, extended version of the Protein Topology Graph Library

web server. The Protein Topology Graph Library describes the protein topology on

the super-secondary structure level. It allows to compute and visualize protein

ligand graphs and search for protein structural motifs. The new server features

additional information on ligand binding to secondary structure elements,

increased usability and an application programming interface (API) to retrieve

data, allowing for an automated analysis of protein topology.

AVAILABILITY AND IMPLEMENTATION: The Protein Topology Graph Library server is

freely available on the web at http://ptgl.uni-frankfurt.de. The website is

implemented in PHP, JavaScript, PostgreSQL and Apache. It is supported by all

major browsers. The VPLG software that was used to compute the protein ligand

graphs and all other data in the database is available under the GNU public

license 2.0 from http://vplg.sourceforge.net.

CONTACT: tim.schaefer@bioinformatik.uni-frankfurt.de;

ina.koch@bioinformatik.uni-frankfurt.de

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv574

PMID: 26446136 [Indexed for MEDLINE]

351. J Ultrasound Med. 2016 Feb;35(2):435-9. doi: 10.7863/ultra.15.02024. Epub 2016

Jan 13.

M.mode.ify: A Free Online Tool to Generate Post Hoc M-Mode Images From Any

Ultrasound Clip.

Smith BC 3rd(1), Avila J(2).

Author information:

(1)Department of Emergency Medicine, University of Tennessee College of Medicine,

Chattanooga, Tennessee USA bensmith.md@gmail.com. (2)Department of Emergency

Medicine, University of Tennessee College of Medicine, Chattanooga, Tennessee

USA.

We present a software tool designed to generate an M-mode image post hoc from any

B-mode ultrasound clip, along any possible axis. M.mode.ify works by breaking

down an ultrasound clip into individual frames. It then rotates and crops these

frames by using a user-selected M-mode line. The post hoc M-mode image is created

by splicing these frames together. Users can measure time and distance after

proper calibration through the M.mode.ify interface. This tool opens up new

possibilities for clinical application, quality assurance, and research. It is

available free for public use at http://www.ultrasoundoftheweek.com/M.mode.ify/.

© 2016 by the American Institute of Ultrasound in Medicine.

DOI: 10.7863/ultra.15.02024

PMID: 26764277 [Indexed for MEDLINE]

352. Mol Biosyst. 2016 Feb;12(2):490-8. doi: 10.1039/c5mb00681c.

HydPred: a novel method for the identification of protein hydroxylation sites

that reveals new insights into human inherited disease.

Li S(1), Lu J(2), Li J(3), Chen X(4), Yao X(1), Xi L(5).

Author information:

(1)College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou,

730000, China. (2)School of Basic Medical Sciences, Lanzhou University, 730000,

China. (3)School of Pharmacy, Lanzhou University, Lanzhou, 730000, China. (4)Key

Laboratory of Desert and Desertification, Cold and Arid Regions Environmental and

Engineering Research Institute, Chinese Academy of Sciences, 730000, China.

(5)Department of Pharmacy, First Hospital of Lanzhou University, Lanzhou, 730000,

China. xill@lzu.edu.cn.

The disruption of protein hydroxylation is highly associated with several serious

diseases and consequently the identification of protein hydroxylation sites has

attracted significant attention recently. Here, we report the development of an

improved method, called HydPred, to identify protein hydroxylation sites

(hydroxyproline and hydroxylysine) based on the synthetic minority over-sampling

technique (SMOTE), the random forest (RF) algorithm and four blocks of newly

composed features that are derived from the protein primary sequence. The HydPred

method achieved the best prediction performance reported until now with Matthew's

correlation coefficient values of 0.770 and 0.857 for hydroxyproline and

hydroxylysine, respectively, according to jack-knife cross-validation. This

represents an improvement of 8% for hydroxyproline and 19% for hydroxylysine

compared to the best results of available predictors. The prediction performance

of HydPred for the external validation of hydroxyproline and hydroxylysine was

also improved compared with other published methods. We subsequently applied

HydPred to study the association of disruption of hydroxylation sites with human

inherited disease. The analyses suggested that the loss of hydroxylation sites is

more likely to cause disease instead of the gain of hydroxylation sites and 52

different human inherited diseases were found to be highly associated with the

loss of hydroxylation sites. Therefore, HydPred represents a new strategy to

discover the molecular basis of pathogenesis associated with abnormal

hydroxylation. HydPred is now available online as a user-friendly web server at

http://lishuyan.lzu.edu.cn/hydpred/.

DOI: 10.1039/c5mb00681c

PMID: 26661679 [Indexed for MEDLINE]

353. Mol Genet Genomics. 2016 Feb;291(1):473-81. doi: 10.1007/s00438-015-1078-7. Epub

2015 Jun 18.

repRNA: a web server for generating various feature vectors of RNA sequences.

Liu B(1,)(2,)(3), Liu F(4), Fang L(5), Wang X(6,)(7), Chou KC(8,)(9).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, 518055, Guangdong, China.

bliu@gordonlifescience.org. (2)Key Laboratory of Network Oriented Intelligent

Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen,

518055, Guangdong, China. bliu@gordonlifescience.org. (3)Gordon Life Science

Institute, Boston, MA, 02478, USA. bliu@gordonlifescience.org. (4)School of

Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate

School, Shenzhen, 518055, Guangdong, China. liufule12@gmail.com. (5)School of

Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate

School, Shenzhen, 518055, Guangdong, China. dragoncloudest@gmail.com. (6)School

of Computer Science and Technology, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, 518055, Guangdong, China. wangxl@insun.hit.edu.cn.

(7)Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute

of Technology Shenzhen Graduate School, Shenzhen, 518055, Guangdong, China.

wangxl@insun.hit.edu.cn. (8)Gordon Life Science Institute, Boston, MA, 02478,

USA. kcchou@gordonlifescience.org. (9)Center of Excellence in Genomic Medicine

Research (CEGMR), King Abdulaziz University, Jeddah, 21589, Saudi Arabia.

kcchou@gordonlifescience.org.

With the rapid growth of RNA sequences generated in the postgenomic age, it is

highly desired to develop a flexible method that can generate various kinds of

vectors to represent these sequences by focusing on their different features.

This is because nearly all the existing machine-learning methods, such as SVM

(support vector machine) and KNN (k-nearest neighbor), can only handle vectors

but not sequences. To meet the increasing demands and speed up the genome

analyses, we have developed a new web server, called "representations of RNA

sequences" (repRNA). Compared with the existing methods, repRNA is much more

comprehensive, flexible and powerful, as reflected by the following facts: (1) it

can generate 11 different modes of feature vectors for users to choose according

to their investigation purposes; (2) it allows users to select the features from

22 built-in physicochemical properties and even those defined by users' own; (3)

the resultant feature vectors and the secondary structures of the corresponding

RNA sequences can be visualized. The repRNA web server is freely accessible to

the public at http://bioinformatics.hitsz.edu.cn/repRNA/ .

DOI: 10.1007/s00438-015-1078-7

PMID: 26085220 [Indexed for MEDLINE]

354. Biol Direct. 2016 Jan 25;11(1):4. doi: 10.1186/s13062-016-0106-9.

BLAST-based structural annotation of protein residues using Protein Data Bank.

Singh H(1), Raghava GP(2).

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A,

Chandigarh, India. harinder@imtech.res.in. (2)Bioinformatics Centre, Institute of

Microbial Technology, Sector 39-A, Chandigarh, India. raghava@imtech.res.in.

BACKGROUND: In the era of next-generation sequencing where thousands of genomes

have been already sequenced; size of protein databases is growing with

exponential rate. Structural annotation of these proteins is one of the biggest

challenges for the computational biologist. Although, it is easy to perform BLAST

search against Protein Data Bank (PDB) but it is difficult for a biologist to

annotate protein residues from BLAST search.

RESULTS: A web-server StarPDB has been developed for structural annotation of a

protein based on its similarity with known protein structures. It uses standard

BLAST software for performing similarity search of a query protein against

protein structures in PDB. This server integrates wide range modules for

assigning different types of annotation that includes, Secondary-structure,

Accessible surface area, Tight-turns, DNA-RNA and Ligand modules. Secondary

structure module allows users to predict regular secondary structure states to

each residue in a protein. Accessible surface area predict the exposed or buried

residues in a protein. Tight-turns module is designed to predict tight turns like

beta-turns in a protein. DNA-RNA module developed for predicting DNA and RNA

interacting residues in a protein. Similarly, Ligand module of server allows one

to predicted ligands, metal and nucleotides ligand interacting residues in a

protein.

CONCLUSIONS: In summary, this manuscript presents a web server for comprehensive

annotation of a protein based on similarity search. It integrates number of

visualization tools that facilitate users to understand structure and function of

protein residues. This web server is available freely for scientific community

from URL http://crdd.osdd.net/raghava/starpdb .

DOI: 10.1186/s13062-016-0106-9

PMCID: PMC4727276

PMID: 26810894 [Indexed for MEDLINE]

355. BioData Min. 2016 Jan 22;9:4. doi: 10.1186/s13040-016-0086-4. eCollection 2016.

Prediction of donor splice sites using random forest with a new sequence encoding

approach.

Meher PK(1), Sahu TK(2), Rao AR(2).

Author information:

(1)Division of Statistical Genetics, Indian Agricultural Statistics Research

Institute, New Delhi, 110 012 India. (2)Centre for Agricultural Bioinformatics,

Indian Agricultural Statistics Research Institute, New Delhi, 110 012 India.

BACKGROUND: Detection of splice sites plays a key role for predicting the gene

structure and thus development of efficient analytical methods for splice site

prediction is vital. This paper presents a novel sequence encoding approach based

on the adjacent di-nucleotide dependencies in which the donor splice site motifs

are encoded into numeric vectors. The encoded vectors are then used as input in

Random Forest (RF), Support Vector Machines (SVM) and Artificial Neural Network

(ANN), Bagging, Boosting, Logistic regression, kNN and Naïve Bayes classifiers

for prediction of donor splice sites.

RESULTS: The performance of the proposed approach is evaluated on the donor

splice site sequence data of Homo sapiens, collected from Homo Sapiens Splice

Sites Dataset (HS3D). The results showed that RF outperformed all the considered

classifiers. Besides, RF achieved higher prediction accuracy than the existing

methods viz., MEM, MDD, WMM, MM1, NNSplice and SpliceView, while compared using

an independent test dataset.

CONCLUSION: Based on the proposed approach, we have developed an online

prediction server (MaLDoSS) to help the biological community in predicting the

donor splice sites. The server is made freely available at

http://cabgrid.res.in:8080/maldoss. Due to computational feasibility and high

prediction accuracy, the proposed approach is believed to help in predicting the

eukaryotic gene structure.

DOI: 10.1186/s13040-016-0086-4

PMCID: PMC4724119

PMID: 26807151

356. BMC Bioinformatics. 2016 Jan 20;17:43. doi: 10.1186/s12859-016-0887-y.

Protein Sequence Annotation Tool (PSAT): a centralized web-based meta-server for

high-throughput sequence annotations.

Leung E(1,)(2), Huang A(3), Cadag E(4,)(5), Montana A(6,)(7), Soliman JL(8,)(9),

Zhou CL(10).

Author information:

(1)Computing Applications and Research, Global Security Computing Applications

Division, Lawrence Livermore National Security, Livermore, CA, 94550, USA.

elo.leung@personalis.com. (2)Personalis, Menlo Park, CA, 94025, USA.

elo.leung@personalis.com. (3)Computing Applications and Research, Global Security

Computing Applications Division, Lawrence Livermore National Security, Livermore,

CA, 94550, USA. huang30@llnl.gov. (4)Computing Applications and Research, Global

Security Computing Applications Division, Lawrence Livermore National Security,

Livermore, CA, 94550, USA. eithon@capellabio.com. (5)Capella Biosciences, Palo

Alto, CA, USA. eithon@capellabio.com. (6)Computing Applications and Research,

Global Security Computing Applications Division, Lawrence Livermore National

Security, Livermore, CA, 94550, USA. aldrin.montana@personalis.com.

(7)Personalis, Menlo Park, CA, 94025, USA. aldrin.montana@personalis.com.

(8)Computing Applications and Research, Global Security Computing Applications

Division, Lawrence Livermore National Security, Livermore, CA, 94550, USA.

jan.v.soliman@gmail.com. (9)LinkedIn, Mountain View, CA, 94043, USA.

jan.v.soliman@gmail.com. (10)Computing Applications and Research, Global Security

Computing Applications Division, Lawrence Livermore National Security, Livermore,

CA, 94550, USA. zhou4@llnl.gov.

BACKGROUND: Here we introduce the Protein Sequence Annotation Tool (PSAT), a

web-based, sequence annotation meta-server for performing integrated,

high-throughput, genome-wide sequence analyses. Our goals in building PSAT were

to (1) create an extensible platform for integration of multiple sequence-based

bioinformatics tools, (2) enable functional annotations and enzyme predictions

over large input protein fasta data sets, and (3) provide a web interface for

convenient execution of the tools.

RESULTS: In this paper, we demonstrate the utility of PSAT by annotating the

predicted peptide gene products of Herbaspirillum sp. strain RV1423, importing

the results of PSAT into EC2KEGG, and using the resulting functional comparisons

to identify a putative catabolic pathway, thereby distinguishing RV1423 from a

well annotated Herbaspirillum species. This analysis demonstrates that

high-throughput enzyme predictions, provided by PSAT processing, can be used to

identify metabolic potential in an otherwise poorly annotated genome.

CONCLUSIONS: PSAT is a meta server that combines the results from several

sequence-based annotation and function prediction codes, and is available at

http://psat.llnl.gov/psat/. PSAT stands apart from other sequence-based genome

annotation systems in providing a high-throughput platform for rapid de novo

enzyme predictions and sequence annotations over large input protein sequence

data sets in FASTA. PSAT is most appropriately applied in annotation of large

protein FASTA sets that may or may not be associated with a single genome.

DOI: 10.1186/s12859-016-0887-y

PMCID: PMC4721133

PMID: 26792120 [Indexed for MEDLINE]

357. PLoS One. 2016 Jan 20;11(1):e0147097. doi: 10.1371/journal.pone.0147097.

eCollection 2016.

CHSalign: A Web Server That Builds upon Junction-Explorer and RNAJAG for Pairwise

Alignment of RNA Secondary Structures with Coaxial Helical Stacking.

Hua L(1), Song Y(1), Kim N(2), Laing C(1), Wang JT(1), Schlick T(2,)(3).

Author information:

(1)Bioinformatics Laboratory, Department of Computer Science, New Jersey

Institute of Technology, Newark, New Jersey, United States of America.

(2)Department of Chemistry, New York University, New York, New York, United

States of America. (3)Courant Institute of Mathematical Sciences, New York

University, New York, New York, United States of America.

RNA junctions are important structural elements of RNA molecules. They are formed

when three or more helices come together in three-dimensional space. Recent

studies have focused on the annotation and prediction of coaxial helical stacking

(CHS) motifs within junctions. Here we exploit such predictions to develop an

efficient alignment tool to handle RNA secondary structures with CHS motifs.

Specifically, we build upon our Junction-Explorer software for predicting coaxial

stacking and RNAJAG for modelling junction topologies as tree graphs to

incorporate constrained tree matching and dynamic programming algorithms into a

new method, called CHSalign, for aligning the secondary structures of RNA

molecules containing CHS motifs. Thus, CHSalign is intended to be an efficient

alignment tool for RNAs containing similar junctions. Experimental results based

on thousands of alignments demonstrate that CHSalign can align two RNA secondary

structures containing CHS motifs more accurately than other RNA secondary

structure alignment tools. CHSalign yields a high score when aligning two RNA

secondary structures with similar CHS motifs or helical arrangement patterns, and

a low score otherwise. This new method has been implemented in a web server, and

the program is also made freely available, at

http://bioinformatics.njit.edu/CHSalign/.

DOI: 10.1371/journal.pone.0147097

PMCID: PMC4720362

PMID: 26789998 [Indexed for MEDLINE]

358. Molecules. 2016 Jan 19;21(1):E95. doi: 10.3390/molecules21010095.

iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein

Binding Sites by Optimizing Imbalanced Training Datasets.

Jia J(1,)(2), Liu Z(3), Xiao X(4,)(5), Liu B(6), Chou KC(7,)(8).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China. jjia@gordonlifescience.org. (2)Gordon Life Science Institute, Boston, MA

02478, USA. jjia@gordonlifescience.org. (3)Computer Department, Jing-De-Zhen

Ceramic Institute, Jing-De-Zhen 333403, China. liuzi189836@163.com. (4)Computer

Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China.

xxiao@gordonlifescience.org. (5)Gordon Life Science Institute, Boston, MA 02478,

USA. xxiao@gordonlifescience.org. (6)Computer Department, Jing-De-Zhen Ceramic

Institute, Jing-De-Zhen 333403, China. lbx1966@163.com. (7)Gordon Life Science

Institute, Boston, MA 02478, USA. kcchou@gordonlifescience.org. (8)Center of

Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University,

Jeddah 21589, Saudi Arabia. kcchou@gordonlifescience.org.

Knowledge of protein-protein interactions and their binding sites is

indispensable for in-depth understanding of the networks in living cells. With

the avalanche of protein sequences generated in the postgenomic age, it is

critical to develop computational methods for identifying in a timely fashion the

protein-protein binding sites (PPBSs) based on the sequence information alone

because the information obtained by this way can be used for both biomedical

research and drug development. To address such a challenge, we have proposed a

new predictor, called iPPBS-Opt, in which we have used: (1) the K-Nearest

Neighbors Cleaning (KNNC) and Inserting Hypothetical Training Samples (IHTS)

treatments to optimize the training dataset; (2) the ensemble voting approach to

select the most relevant features; and (3) the stationary wavelet transform to

formulate the statistical samples. Cross-validation tests by targeting the

experiment-confirmed results have demonstrated that the new predictor is very

promising, implying that the aforementioned practices are indeed very effective.

Particularly, the approach of using the wavelets to express protein/peptide

sequences might be the key in grasping the problem's essence, fully consistent

with the findings that many important biological functions of proteins can be

elucidated with their low-frequency internal motions. To maximize the convenience

of most experimental scientists, we have provided a step-by-step guide on how to

use the predictor's web server (http://www.jci-bioinfo.cn/iPPBS-Opt) to get the

desired results without the need to go through the complicated mathematical

equations involved.

DOI: 10.3390/molecules21010095

PMID: 26797600 [Indexed for MEDLINE]

359. Bioinformatics. 2016 Jan 15;32(2):309-11. doi: 10.1093/bioinformatics/btv557.

Epub 2015 Sep 28.

Cytoscape.js: a graph theory library for visualisation and analysis.

Franz M(1), Lopes CT(1), Huck G(1), Dong Y(1), Sumer O(1), Bader GD(1).

Author information:

(1)The Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada.

Cytoscape.js is an open-source JavaScript-based graph library. Its most common

use case is as a visualization software component, so it can be used to render

interactive graphs in a web browser. It also can be used in a headless manner,

useful for graph operations on a server, such as Node.js.AVAILABILITY AND

IMPLEMENTATION: Cytoscape.js is implemented in JavaScript. Documentation,

downloads and source code are available at http://js.cytoscape.org.

CONTACT: gary.bader@utoronto.ca.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv557

PMCID: PMC4708103

PMID: 26415722 [Indexed for MEDLINE]

360. Methods. 2016 Jan 15;93:15-23. doi: 10.1016/j.ymeth.2015.08.021. Epub 2015 Aug

28.

Enhancing protein function prediction with taxonomic constraints--The Argot2.5

web server.

Lavezzo E(1), Falda M(1), Fontana P(2), Bianco L(2), Toppo S(3).

Author information:

(1)Department of Molecular Medicine, University of Padova, Padova, Italy.

(2)Istituto Agrario San Michele all'Adige Research and Innovation Centre,

Foundation Edmund Mach, Trento, Italy. (3)Department of Molecular Medicine,

University of Padova, Padova, Italy. Electronic address: stefano.toppo@unipd.it.

Argot2.5 (Annotation Retrieval of Gene Ontology Terms) is a web server designed

to predict protein function. It is an updated version of the previous Argot2

enriched with new features in order to enhance its usability and its overall

performance. The algorithmic strategy exploits the grouping of Gene Ontology

terms by means of semantic similarity to infer protein function. The tool has

been challenged over two independent benchmarks and compared to Argot2, PANNZER,

and a baseline method relying on BLAST, proving to obtain a better performance

thanks to the contribution of some key interventions in critical steps of the

working pipeline. The most effective changes regard: (a) the selection of the

input data from sequence similarity searches performed against a clustered

version of UniProt databank and a remodeling of the weights given to Pfam hits,

(b) the application of taxonomic constraints to filter out annotations that

cannot be applied to proteins belonging to the species under investigation. The

taxonomic rules are derived from our in-house developed tool, FunTaxIS, that

extends those provided by the Gene Ontology consortium. The web server is free

for academic users and is available online at

http://www.medcomp.medicina.unipd.it/Argot2-5/.

Copyright © 2015 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ymeth.2015.08.021

PMID: 26318087 [Indexed for MEDLINE]

361. Methods. 2016 Jan 15;93:72-83. doi: 10.1016/j.ymeth.2015.07.004. Epub 2015 Jul

10.

Modeling of protein-peptide interactions using the CABS-dock web server for

binding site search and flexible docking.

Blaszczyk M(1), Kurcinski M(1), Kouza M(1), Wieteska L(1), Debinski A(1),

Kolinski A(1), Kmiecik S(2).

Author information:

(1)Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw,

Poland. (2)Department of Chemistry, University of Warsaw, Pasteura 1, 02-093

Warsaw, Poland. Electronic address: sekmi@chem.uw.edu.pl.

Protein-peptide interactions play essential functional roles in living organisms

and their structural characterization is a hot subject of current experimental

and theoretical research. Computational modeling of the structure of

protein-peptide interactions is usually divided into two stages: prediction of

the binding site at a protein receptor surface, and then docking (and modeling)

the peptide structure into the known binding site. This paper presents a

comprehensive CABS-dock method for the simultaneous search of binding sites and

flexible protein-peptide docking, available as a user's friendly web server. We

present example CABS-dock results obtained in the default CABS-dock mode and

using its advanced options that enable the user to increase the range of

flexibility for chosen receptor fragments or to exclude user-selected binding

modes from docking search. Furthermore, we demonstrate a strategy to improve

CABS-dock performance by assessing the quality of models with classical molecular

dynamics. Finally, we discuss the promising extensions and applications of the

CABS-dock method and provide a tutorial appendix for the convenient analysis and

visualization of CABS-dock results. The CABS-dock web server is freely available

at http://biocomp.chem.uw.edu.pl/CABSdock/.

Copyright © 2016 The Authors. Published by Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ymeth.2015.07.004

PMID: 26165956 [Indexed for MEDLINE]

362. BMC Genomics. 2016 Jan 13;17:49. doi: 10.1186/s12864-015-2346-y.

CANEapp: a user-friendly application for automated next generation transcriptomic

data analysis.

Velmeshev D(1,)(2), Lally P(3,)(4), Magistri M(5), Faghihi MA(6).

Author information:

(1)Department of Psychiatry, University of Miami Miller School of Medicine,

Miami, FL, 33136, USA. dvelmeshev@med.miami.edu. (2)Department of Biochemistry &

Molecular Biology, University of Miami Miller School of Medicine, Miami, FL,

33136, USA. dvelmeshev@med.miami.edu. (3)Department of Psychiatry, University of

Miami Miller School of Medicine, Miami, FL, 33136, USA. p.lally1@umiami.edu.

(4)Department of Biomedical Engineering, University of Miami, Coral Gables, FL,

33146, USA. p.lally1@umiami.edu. (5)Department of Psychiatry, University of Miami

Miller School of Medicine, Miami, FL, 33136, USA. mmagistri@med.miami.edu.

(6)Department of Psychiatry, University of Miami Miller School of Medicine,

Miami, FL, 33136, USA. MFaghihi@med.miami.edu.

BACKGROUND: Next generation sequencing (NGS) technologies are indispensable for

molecular biology research, but data analysis represents the bottleneck in their

application. Users need to be familiar with computer terminal commands, the Linux

environment, and various software tools and scripts. Analysis workflows have to

be optimized and experimentally validated to extract biologically meaningful

data. Moreover, as larger datasets are being generated, their analysis requires

use of high-performance servers.

RESULTS: To address these needs, we developed CANEapp (application for

Comprehensive automated Analysis of Next-generation sequencing Experiments), a

unique suite that combines a Graphical User Interface (GUI) and an automated

server-side analysis pipeline that is platform-independent, making it suitable

for any server architecture. The GUI runs on a PC or Mac and seamlessly connects

to the server to provide full GUI control of RNA-sequencing (RNA-seq) project

analysis. The server-side analysis pipeline contains a framework that is

implemented on a Linux server through completely automated installation of

software components and reference files. Analysis with CANEapp is also fully

automated and performs differential gene expression analysis and novel noncoding

RNA discovery through alternative workflows (Cuffdiff and R packages edgeR and

DESeq2). We compared CANEapp to other similar tools, and it significantly

improves on previous developments. We experimentally validated CANEapp's

performance by applying it to data derived from different experimental paradigms

and confirming the results with quantitative real-time PCR (qRT-PCR). CANEapp

adapts to any server architecture by effectively using available resources and

thus handles large amounts of data efficiently. CANEapp performance has been

experimentally validated on various biological datasets. CANEapp is available

free of charge at

http://psychiatry.med.miami.edu/research/laboratory-of-translational-rna-genomics

/CANE-app .

CONCLUSIONS: We believe that CANEapp will serve both biologists with no

computational experience and bioinformaticians as a simple, timesaving but

accurate and powerful tool to analyze large RNA-seq datasets and will provide

foundations for future development of integrated and automated high-throughput

genomics data analysis tools. Due to its inherently standardized pipeline and

combination of automated analysis and platform-independence, CANEapp is an ideal

for large-scale collaborative RNA-seq projects between different institutions and

research groups.

DOI: 10.1186/s12864-015-2346-y

PMCID: PMC4710974

PMID: 26758513 [Indexed for MEDLINE]

363. Sci Rep. 2016 Jan 12;6:19062. doi: 10.1038/srep19062.

iMiRNA-SSF: Improving the Identification of MicroRNA Precursors by Combining

Negative Sets with Different Distributions.

Chen J(1), Wang X(1,)(2), Liu B(1,)(2).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong, China. (2)Key Laboratory of

Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, Guangdong, China.

The identification of microRNA precursors (pre-miRNAs) helps in understanding

regulator in biological processes. The performance of computational predictors

depends on their training sets, in which the negative sets play an important

role. In this regard, we investigated the influence of benchmark datasets on the

predictive performance of computational predictors in the field of miRNA

identification, and found that the negative samples have significant impact on

the predictive results of various methods. We constructed a new benchmark set

with different data distributions of negative samples. Trained with this high

quality benchmark dataset, a new computational predictor called iMiRNA-SSF was

proposed, which employed various features extracted from RNA sequences.

Experimental results showed that iMiRNA-SSF outperforms three state-of-the-art

computational methods. For practical applications, a web-server of iMiRNA-SSF was

established at the website http://bioinformatics.hitsz.edu.cn/iMiRNA-SSF/.

DOI: 10.1038/srep19062

PMCID: PMC4709562

PMID: 26753561 [Indexed for MEDLINE]

364. Sci Rep. 2016 Jan 11;6:19016. doi: 10.1038/srep19016.

Rsite2: an efficient computational method to predict the functional sites of

noncoding RNAs.

Zeng P(1), Cui Q(1).

Author information:

(1)Department of Biomedical Informatics, Centre for Noncoding RNA Medicine,

School of Basic Medical Sciences, Peking University, 38 Xueyuan Rd, Beijing,

100191, China.

Noncoding RNAs (ncRNAs) represent a big class of important RNA molecules. Given

the large number of ncRNAs, identifying their functional sites is becoming one of

the most important topics in the post-genomic era, but available computational

methods are limited. For the above purpose, we previously presented a tertiary

structure based method, Rsite, which first calculates the distance metrics

defined in Methods with the tertiary structure of an ncRNA and then identifies

the nucleotides located within the extreme points in the distance curve as the

functional sites of the given ncRNA. However, the application of Rsite is largely

limited because of limited RNA tertiary structures. Here we present a secondary

structure based computational method, Rsite2, based on the observation that the

secondary structure based nucleotide distance is strongly positively correlated

with that derived from tertiary structure. This makes it reasonable to replace

tertiary structure with secondary structure, which is much easier to obtain and

process. Moreover, we applied Rsite2 to three ncRNAs (tRNA (Lys), Diels-Alder

ribozyme, and RNase P) and a list of human mitochondria transcripts. The results

show that Rsite2 works well with nearly equivalent accuracy as Rsite but is much

more feasible and efficient. Finally, a web-server, the source codes, and the

dataset of Rsite2 are available at http://www.cuialb.cn/rsite2.

DOI: 10.1038/srep19016

PMCID: PMC4707467

PMID: 26751501 [Indexed for MEDLINE]

365. Front Plant Sci. 2016 Jan 5;6:1194. doi: 10.3389/fpls.2015.01194. eCollection

2015.

CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization

Tool.

Tzfadia O(1), Diels T(1), De Meyer S(2), Vandepoele K(1), Aharoni A(3), Van de

Peer Y(4).

Author information:

(1)Department of Plant Systems Biology, Vlaams Instituut voor

BiotechnologieGhent, Belgium; Department of Plant Biotechnology and

Bioinformatics, Ghent UniversityGhent, Belgium; Bioinformatics Institute Ghent,

Ghent UniversityGhent, Belgium. (2)Department of Plant Systems Biology, Vlaams

Instituut voor BiotechnologieGhent, Belgium; Department of Plant Biotechnology

and Bioinformatics, Ghent UniversityGhent, Belgium. (3)Department of Plant

Sciences and the Environment, Weizmann Institute of Science Rehovot, Israel.

(4)Department of Plant Systems Biology, Vlaams Instituut voor

BiotechnologieGhent, Belgium; Department of Plant Biotechnology and

Bioinformatics, Ghent UniversityGhent, Belgium; Bioinformatics Institute Ghent,

Ghent UniversityGhent, Belgium; Genomics Research Institute, University of

PretoriaPretoria, South Africa.

MOTIVATION: Comparative transcriptomics is a common approach in functional gene

discovery efforts. It allows for finding conserved co-expression patterns between

orthologous genes in closely related plant species, suggesting that these genes

potentially share similar function and regulation. Several efficient

co-expression-based tools have been commonly used in plant research but most of

these pipelines are limited to data from model systems, which greatly limit their

utility. Moreover, in addition, none of the existing pipelines allow plant

researchers to make use of their own unpublished gene expression data for

performing a comparative co-expression analysis and generate multi-species

co-expression networks.

RESULTS: We introduce CoExpNetViz, a computational tool that uses a set of query

or "bait" genes as an input (chosen by the user) and a minimum of one

pre-processed gene expression dataset. The CoExpNetViz algorithm proceeds in

three main steps; (i) for every bait gene submitted, co-expression values are

calculated using mutual information and Pearson correlation coefficients, (ii)

non-bait (or target) genes are grouped based on cross-species orthology, and

(iii) output files are generated and results can be visualized as network graphs

in Cytoscape.

AVAILABILITY: The CoExpNetViz tool is freely available both as a PHP web server

(link: http://bioinformatics.psb.ugent.be/webtools/coexpr/) (implemented in C++)

and as a Cytoscape plugin (implemented in Java). Both versions of the CoExpNetViz

tool support LINUX and Windows platforms.

DOI: 10.3389/fpls.2015.01194

PMCID: PMC4700130

PMID: 26779228

366. PLoS One. 2016 Jan 5;11(1):e0146409. doi: 10.1371/journal.pone.0146409.

eCollection 2016.

SeqFeatR for the Discovery of Feature-Sequence Associations.

Budeus B(1), Timm J(2), Hoffmann D(1).

Author information:

(1)Research Group Bioinformatics, Faculty of Biology, University of

Duisburg-Essen, Essen, NRW, Germany. (2)Institute for Virology, University

Hospital Düsseldorf, Düsseldorf, NRW, Germany.

Specific selection pressures often lead to specifically mutated genomes. The open

source software SeqFeatR has been developed to identify associations between

mutation patterns in biological sequences and specific selection pressures

("features"). For instance, SeqFeatR has been used to discover in viral protein

sequences new T cell epitopes for hosts of given HLA types. SeqFeatR supports

frequentist and Bayesian methods for the discovery of statistical

sequence-feature associations. Moreover, it offers novel ways to visualize

results of the statistical analyses and to relate them to further properties. In

this article we demonstrate various functions of SeqFeatR with real data. The

most frequently used set of functions is also provided by a web server. SeqFeatR

is implemented as R package and freely available from the R archive CRAN

(http://cran.r-project.org/web/packages/SeqFeatR/index.html). The package

includes a tutorial vignette. The software is distributed under the GNU General

Public License (version 3 or later). The web server URL is

https://seqfeatr.zmb.uni-due.de.

DOI: 10.1371/journal.pone.0146409

PMCID: PMC4701496

PMID: 26731669 [Indexed for MEDLINE]

367. Nucleic Acids Res. 2016 Jan 4;44(D1):D196-202. doi: 10.1093/nar/gkv1273. Epub

2015 Nov 20.

deepBase v2.0: identification, expression, evolution and function of small RNAs,

LncRNAs and circular RNAs from deep-sequencing data.

Zheng LL(1), Li JH(1), Wu J(1), Sun WJ(1), Liu S(1), Wang ZL(1), Zhou H(1), Yang

JH(2), Qu LH(2).

Author information:

(1)RNA Information Center, Key Laboratory of Gene Engineering of the Ministry of

Education, State Key Laboratory for Biocontrol, Sun Yat-sen University, Guangzhou

510275, P.R. China. (2)RNA Information Center, Key Laboratory of Gene Engineering

of the Ministry of Education, State Key Laboratory for Biocontrol, Sun Yat-sen

University, Guangzhou 510275, P.R. China lssqlh@mail.sysu.edu.cn.

Small non-coding RNAs (e.g. miRNAs) and long non-coding RNAs (e.g. lincRNAs and

circRNAs) are emerging as key regulators of various cellular processes. However,

only a very small fraction of these enigmatic RNAs have been well functionally

characterized. In this study, we describe deepBase v2.0

(http://biocenter.sysu.edu.cn/deepBase/), an updated platform, to decode

evolution, expression patterns and functions of diverse ncRNAs across 19 species.

deepBase v2.0 has been updated to provide the most comprehensive collection of

ncRNA-derived small RNAs generated from 588 sRNA-Seq datasets. Moreover, we

developed a pipeline named lncSeeker to identify 176 680 high-confidence lncRNAs

from 14 species. Temporal and spatial expression patterns of various ncRNAs were

profiled. We identified approximately 24 280 primate-specific, 5193

rodent-specific lncRNAs, and 55 highly conserved lncRNA orthologs between human

and zebrafish. We annotated 14 867 human circRNAs, 1260 of which are orthologous

to mouse circRNAs. By combining expression profiles and functional genomic

annotations, we developed lncFunction web-server to predict the function of

lncRNAs based on protein-lncRNA co-expression networks. This study is expected to

provide considerable resources to facilitate future experimental studies and to

uncover ncRNA functions.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv1273

PMCID: PMC4702900

PMID: 26590255 [Indexed for MEDLINE]

368. Nucleic Acids Res. 2016 Jan 4;44(D1):D801-7. doi: 10.1093/nar/gkv1204. Epub 2015

Nov 17.

InsectBase: a resource for insect genomes and transcriptomes.

Yin C(1), Shen G(2), Guo D(1), Wang S(3), Ma X(4), Xiao H(5), Liu J(6), Zhang

Z(7), Liu Y(8), Zhang Y(6), Yu K(4), Huang S(6), Li F(9).

Author information:

(1)Ministry of Agriculture, Key Lab of Agricultural Entomology and Institute of

Insect Sciences, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, China

Department of Entomology, Nanjing Agricultural University, Nanjing 210095, China.

(2)Library of Nanjing Agricultural University, Nanjing 210095, China.

(3)Technical Center for Animal Plant and Food Inspection and Quarantine, Shanghai

Entry-Exit Inspection and Quarantine Bureau, Shanghai 200135, China.

(4)Department of Entomology, Nanjing Agricultural University, Nanjing 210095,

China. (5)Department of Entomology, Nanjing Agricultural University, Nanjing

210095, China Department of City Construction, Shaoyang University, Shaoyang

422000, China. (6)College of Information Science and Technology, Nanjing

Agricultural University, Nanjing, 210095 Jiangsu, China. (7)Department of

Entomology, Nanjing Agricultural University, Nanjing 210095, China Hubei Insect

Resources Utilization and Sustainable Pest Management Key Laboratory, College of

Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070,

Hubei, China. (8)Department of Entomology, Nanjing Agricultural University,

Nanjing 210095, China Department of Entomology, China Agricultural University,

Beijing 100193, China. (9)Ministry of Agriculture, Key Lab of Agricultural

Entomology and Institute of Insect Sciences, Zhejiang University, 866 Yuhangtang

Road, Hangzhou 310058, China lifei18@zju.edu.cn.

The genomes and transcriptomes of hundreds of insects have been sequenced.

However, insect community lacks an integrated, up-to-date collection of insect

gene data. Here, we introduce the first release of InsectBase, available online

at http://www.insect-genome.com. The database encompasses 138 insect genomes, 116

insect transcriptomes, 61 insect gene sets, 36 gene families of 60 insects, 7544

miRNAs of 69 insects, 96,925 piRNAs of Drosophila melanogaster and Chilo

suppressalis, 2439 lncRNA of Nilaparvata lugens, 22,536 pathways of 78 insects,

678,881 untranslated regions (UTR) of 84 insects and 160,905 coding sequences

(CDS) of 70 insects. This release contains over 12 million sequences and provides

search functionality, a BLAST server, GBrowse, insect pathway construction, a

Facebook-like network for the insect community (iFacebook), and phylogenetic

analysis of selected genes.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv1204

PMCID: PMC4702856

PMID: 26578584 [Indexed for MEDLINE]

369. Nucleic Acids Res. 2016 Jan 4;44(D1):D793-800. doi: 10.1093/nar/gkv1208. Epub

2015 Nov 17.

Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine.

Elsik CG(1), Tayal A(2), Diesh CM(2), Unni DR(2), Emery ML(3), Nguyen HN(4),

Hagen DE(2).

Author information:

(1)Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA

Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA MU

Informatics Institute, University of Missouri, Columbia, MO 65211, USA

elsikc@missouri.edu. (2)Division of Animal Sciences, University of Missouri,

Columbia, MO 65211, USA. (3)Division of Plant Sciences, University of Missouri,

Columbia, MO 65211, USA. (4)MU Informatics Institute, University of Missouri,

Columbia, MO 65211, USA.

We report an update of the Hymenoptera Genome Database (HGD)

(http://HymenopteraGenome.org), a model organism database for insect species of

the order Hymenoptera (ants, bees and wasps). HGD maintains genomic data for 9

bee species, 10 ant species and 1 wasp, including the versions of genome and

annotation data sets published by the genome sequencing consortiums and those

provided by NCBI. A new data-mining warehouse, HymenopteraMine, based on the

InterMine data warehousing system, integrates the genome data with data from

external sources and facilitates cross-species analyses based on orthology. New

genome browsers and annotation tools based on JBrowse/WebApollo provide easy

genome navigation, and viewing of high throughput sequence data sets and can be

used for collaborative genome annotation. All of the genomes and annotation data

sets are combined into a single BLAST server that allows users to select and

combine sequence data sets to search.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv1208

PMCID: PMC4702858

PMID: 26578564 [Indexed for MEDLINE]

370. Nucleic Acids Res. 2016 Jan 4;44(D1):D848-54. doi: 10.1093/nar/gkv1155. Epub 2015

Nov 2.

MouseNet v2: a database of gene networks for studying the laboratory mouse and

eight other model vertebrates.

Kim E(1), Hwang S(2), Kim H(1), Shim H(1), Kang B(1), Yang S(1), Shim JH(1), Shin

SY(1), Marcotte EM(3), Lee I(4).

Author information:

(1)Department of Biotechnology, College of Life Science and Biotechnology, Yonsei

University, Seoul, Korea. (2)Department of Biotechnology, College of Life Science

and Biotechnology, Yonsei University, Seoul, Korea Center for Systems and

Synthetic Biology, Institute for Cellular and Molecular Biology, University of

Texas at Austin, TX 78712, USA. (3)Center for Systems and Synthetic Biology,

Institute for Cellular and Molecular Biology, University of Texas at Austin, TX

78712, USA. (4)Department of Biotechnology, College of Life Science and

Biotechnology, Yonsei University, Seoul, Korea insuklee@yonsei.ac.kr.

Laboratory mouse, Mus musculus, is one of the most important animal tools in

biomedical research. Functional characterization of the mouse genes, hence, has

been a long-standing goal in mammalian and human genetics. Although large-scale

knockout phenotyping is under progress by international collaborative efforts, a

large portion of mouse genome is still poorly characterized for cellular

functions and associations with disease phenotypes. A genome-scale functional

network of mouse genes, MouseNet, was previously developed in context of

MouseFunc competition, which allowed only limited input data for network

inferences. Here, we present an improved mouse co-functional network, MouseNet v2

(available at http://www.inetbio.org/mousenet), which covers 17 714 genes (>88%

of coding genome) with 788 080 links, along with a companion web server for

network-assisted functional hypothesis generation. The network database has been

substantially improved by large expansion of genomics data. For example, MouseNet

v2 database contains 183 co-expression networks inferred from 8154 public

microarray samples. We demonstrated that MouseNet v2 is predictive for mammalian

phenotypes as well as human diseases, which suggests its usefulness in discovery

of novel disease genes and dissection of disease pathways. Furthermore, MouseNet

v2 database provides functional networks for eight other vertebrate models used

in various research fields.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv1155

PMCID: PMC4702832

PMID: 26527726 [Indexed for MEDLINE]

371. Biochim Biophys Acta. 2016 Jan;1864(1):11-9. doi: 10.1016/j.bbapap.2015.10.004.

Epub 2015 Oct 22.

ProTSAV: A protein tertiary structure analysis and validation server.

Singh A(1), Kaushik R(2), Mishra A(2), Shanker A(3), Jayaram B(4).

Author information:

(1)Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi,

India; Department of Bioinformatics, Banasthali Vidyapith, Banasthali, 304022,

India. (2)Supercomputing Facility for Bioinformatics & Computational Biology, IIT

Delhi, India; Kusuma School of Biological Sciences, IIT Delhi, India.

(3)Department of Bioinformatics, Banasthali Vidyapith, Banasthali, 304022, India.

(4)Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi,

India; Kusuma School of Biological Sciences, IIT Delhi, India; Department of

Chemistry, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, 110016,

India. Electronic address: bjayaram@scfbio-iitd.res.in.

Quality assessment of predicted model structures of proteins is as important as

the protein tertiary structure prediction. A highly efficient quality assessment

of predicted model structures directs further research on function. Here we

present a new server ProTSAV, capable of evaluating predicted model structures

based on some popular online servers and standalone tools. ProTSAV furnishes the

user with a single quality score in case of individual protein structure along

with a graphical representation and ranking in case of multiple protein structure

assessment. The server is validated on ~64,446 protein structures including

experimental structures from RCSB and predicted model structures for CASP targets

and from public decoy sets. ProTSAV succeeds in predicting quality of protein

structures with a specificity of 100% and a sensitivity of 98% on experimentally

solved structures and achieves a specificity of 88%and a sensitivity of 91% on

predicted protein structures of CASP11 targets under 2Å.The server overcomes the

limitations of any single server/method and is seen to be robust in helping in

quality assessment. ProTSAV is freely available at

http://www.scfbio-iitd.res.in/software/proteomics/protsav.jsp.

Copyright © 2015 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.bbapap.2015.10.004

PMID: 26478257 [Indexed for MEDLINE]

372. Bioinformatics. 2016 Jan 1;32(1):25-34. doi: 10.1093/bioinformatics/btv525. Epub

2015 Sep 5.

ResiCon: a method for the identification of dynamic domains, hinges and

interfacial regions in proteins.

Dziubiński M(1), Daniluk P(2), Lesyng B(2).

Author information:

(1)Department of Biophysics and CoE BioExploratorium, Faculty of Physics,

University of Warsaw, 02-089 Warsaw, Poland and. (2)Department of Biophysics and

CoE BioExploratorium, Faculty of Physics, University of Warsaw, 02-089 Warsaw,

Poland and Bioinformatics Laboratory, Mossakowski Medical Research Centre, Polish

Academy of Sciences, 02-106 Warsaw, Poland.

MOTIVATION: Structure of most proteins is flexible. Identification and analysis

of intramolecular motions is a complex problem. Breaking a structure into

relatively rigid parts, the so-called dynamic domains, may help comprehend the

complexity of protein's mobility. We propose a new approach called ResiCon

(Residue Contacts analysis), which performs this task by applying a data-mining

analysis of an ensemble of protein configurations and recognizes dynamic domains,

hinges and interfacial regions, by considering contacts between residues.

RESULTS: Dynamic domains found by ResiCon are more compact than those identified

by two other popular methods: PiSQRD and GeoStaS. The current analysis was

carried out using a known reference set of 30 NMR protein structures, as well as

molecular dynamics simulation data of flap opening events in HIV-1 protease. The

more detailed analysis of HIV-1 protease dataset shows that ResiCon identified

dynamic domains involved in structural changes of functional importance.

AVAILABILITY AND IMPLEMENTATION: The ResiCon server is available at URL:

http://dworkowa.imdik.pan.pl/EP/ResiCon.

CONTACT: pawel@bioexploratorium.pl

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv525

PMID: 26342233 [Indexed for MEDLINE]

373. Biosystems. 2016 Jan;139:17-22. doi: 10.1016/j.biosystems.2015.10.004. Epub 2015

Dec 1.

PRIdictor: Protein-RNA Interaction predictor.

Tuvshinjargal N(1), Lee W(1), Park B(1), Han K(2).

Author information:

(1)Department of Computer Science and Engineering, Inha University, Incheon,

South Korea. (2)Department of Computer Science and Engineering, Inha University,

Incheon, South Korea. Electronic address: khan@inha.ac.kr.

Several computational methods have been developed to predict RNA-binding sites in

protein, but its inverse problem (i.e., predicting protein-binding sites in RNA)

has received much less attention. Furthermore, most methods that predict

RNA-binding sites in protein do not consider interaction partners of a protein.

This paper presents a web server called PRIdictor (Protein-RNA Interaction

predictor) which predicts mutual binding sites in RNA and protein at the

nucleotide- and residue-level resolutions from their sequences. PRIdictor can be

used as a web-based application or web service at

http://bclab.inha.ac.kr/pridictor.

Copyright © 2015 Elsevier Ireland Ltd. All rights reserved.

DOI: 10.1016/j.biosystems.2015.10.004

PMID: 26607710 [Indexed for MEDLINE]

374. J Biomol Struct Dyn. 2016;34(1):223-35. doi: 10.1080/07391102.2015.1014422. Epub

2015 Mar 3.

iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair

composition approach.

Liu B(1,)(2,)(3), Fang L(1), Liu F(1), Wang X(1,)(2), Chou KC(3,)(4).

Author information:

(1)a School of Computer Science and Technology , Harbin Institute of Technology

Shenzhen Graduate School , Shenzhen , Guangdong 518055 , China. (2)b Key

Laboratory of Network Oriented Intelligent Computation , Harbin Institute of

Technology Shenzhen Graduate School , Shenzhen , Guangdong , China. (3)c Gordon

Life Science Institute , Belmont , MA 02478 , USA. (4)d Center of Excellence in

Genomic Medicine Research (CEGMR) , King Abdulaziz University , Jeddah , Saudi

Arabia.

A microRNA (miRNA) is a small non-coding RNA molecule, functioning in

transcriptional and post-transcriptional regulation of gene expression. The human

genome may encode over 1000 miRNAs. Albeit poorly characterized, miRNAs are

widely deemed as important regulators of biological processes. Aberrant

expression of miRNAs has been observed in many cancers and other disease states,

indicating that they are deeply implicated with these diseases, particularly in

carcinogenesis. Therefore, it is important for both basic research and

miRNA-based therapy to discriminate the real pre-miRNAs from the false ones (such

as hairpin sequences with similar stem-loops). Particularly, with the avalanche

of RNA sequences generated in the post-genomic age, it is highly desired to

develop computational sequence-based methods for effectively identifying the

human pre-miRNAs. Here, we propose a predictor called "iMiRNA-PseDPC", in which

the RNA sequences are formulated by a novel feature vector called "pseudo

distance-pair composition" (PseDPC) with 10 types of structure statuses. Rigorous

cross-validations on a much larger and more stringent newly constructed benchmark

data-set showed that our approach has remarkably outperformed the existing ones

in either prediction accuracy or efficiency, indicating the new predictor is

quite promising or at least may become a complementary tool to the existing

predictors in this area. For the convenience of most experimental scientists, a

user-friendly web server for the new predictor has been established at

http://bioinformatics.hitsz.edu.cn/iMiRNA-PseDPC/, by which users can easily get

their desired results without the need to go through the mathematical details. It

is anticipated that the new predictor may become a useful high throughput tool

for genome analysis particularly in dealing with large-scale data.

DOI: 10.1080/07391102.2015.1014422

PMID: 25645238 [Indexed for MEDLINE]

375. Methods Mol Biol. 2016;1490:217-35. doi: 10.1007/978-1-4939-6433-8\_14.

RNA 3D Structure Modeling by Combination of Template-Based Method ModeRNA,

Template-Free Folding with SimRNA, and Refinement with QRNAS.

Piatkowski P(1), Kasprzak JM(2,)(3), Kumar D(4), Magnus M(1), Chojnowski G(1),

Bujnicki JM(2,)(3).

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109, Warsaw, Poland.

(2)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109, Warsaw, Poland.

iamb@genesilico.pl. (3)Laboratory of Bioinformatics, Institute of Molecular

Biology and Biotechnology, Adam Mickiewicz University, ul. Umultowska 89, 61-614,

Poznan, Poland. iamb@genesilico.pl. (4)Laboratory of Bioinformatics, Institute of

Molecular Biology and Biotechnology, Adam Mickiewicz University, ul. Umultowska

89, 61-614, Poznan, Poland.

RNA encompasses an essential part of all known forms of life. The functions of

many RNA molecules are dependent on their ability to form complex

three-dimensional (3D) structures. However, experimental determination of RNA 3D

structures is laborious and challenging, and therefore, the majority of known

RNAs remain structurally uncharacterized. To address this problem, computational

structure prediction methods were developed that either utilize information

derived from known structures of other RNA molecules (by way of template-based

modeling) or attempt to simulate the physical process of RNA structure formation

(by way of template-free modeling). All computational methods suffer from various

limitations that make theoretical models less reliable than high-resolution

experimentally determined structures. This chapter provides a protocol for

computational modeling of RNA 3D structure that overcomes major limitations by

combining two complementary approaches: template-based modeling that is capable

of predicting global architectures based on similarity to other molecules but

often fails to predict local unique features, and template-free modeling that can

predict the local folding, but is limited to modeling the structure of relatively

small molecules. Here, we combine the use of a template-based method ModeRNA with

a template-free method SimRNA. ModeRNA requires a sequence alignment of the

target RNA sequence to be modeled with a template of the known structure; it

generates a model that predicts the structure of a conserved core and provides a

starting point for modeling of variable regions. SimRNA can be used to fold small

RNAs (<80 nt) without any additional structural information, and to refold parts

of models for larger RNAs that have a correctly modeled core. ModeRNA can be

either downloaded, compiled and run locally or run through a web interface at

http://genesilico.pl/modernaserver/ . SimRNA is currently available to download

for local use as a precompiled software package at

http://genesilico.pl/software/stand-alone/simrna and as a web server at

http://genesilico.pl/SimRNAweb . For model optimization we use QRNAS, available

at http://genesilico.pl/qrnas .

DOI: 10.1007/978-1-4939-6433-8\_14

PMID: 27665602

376. Methods Mol Biol. 2016;1490:187-98. doi: 10.1007/978-1-4939-6433-8\_12.

Modeling Small Noncanonical RNA Motifs with the Rosetta FARFAR Server.

Yesselman JD(1), Das R(2,)(3).

Author information:

(1)Biochemistry Department, Stanford University, Stanford, CA, 94305, USA.

(2)Biochemistry Department, Stanford University, Stanford, CA, 94305, USA.

rhiju@stanford.edu. (3)Physics Department, Stanford University, Stanford, CA,

94305, USA. rhiju@stanford.edu.

Noncanonical RNA motifs help define the vast complexity of RNA structure and

function, and in many cases, these loops and junctions are on the order of only

ten nucleotides in size. Unfortunately, despite their small size, there is no

reliable method to determine the ensemble of lowest energy structures of

junctions and loops at atomic accuracy. This chapter outlines straightforward

protocols using a webserver for Rosetta Fragment Assembly of RNA with Full Atom

Refinement (FARFAR) ( http://rosie.rosettacommons.org/rna\_denovo/submit ) to

model the 3D structure of small noncanonical RNA motifs for use in visualizing

motifs and for further refinement or filtering with experimental data such as NMR

chemical shifts.

DOI: 10.1007/978-1-4939-6433-8\_12

PMID: 27665600

377. Methods Mol Biol. 2016;1490:73-82. doi: 10.1007/978-1-4939-6433-8\_6.

STarMir Tools for Prediction of microRNA Binding Sites.

Kanoria S(1), Rennie W(1), Liu C(1), Carmack CS(1), Lu J(2), Ding Y(3).

Author information:

(1)Wadsworth Center, New York State Department of Health, Center for Medical

Science, 150 New Scotland Avenue, Albany, NY, 12208, USA. (2)Department of

Genetics and Yale Stem Cell Center, Yale University, New Haven, CT, 06520, USA.

(3)Wadsworth Center, New York State Department of Health, Center for Medical

Science, 150 New Scotland Avenue, Albany, NY, 12208, USA. ye.ding@health.ny.gov.

MicroRNAs (miRNAs) are a class of endogenous short noncoding RNAs that regulate

gene expression by targeting messenger RNAs (mRNAs), which results in

translational repression and/or mRNA degradation. As regulatory molecules, miRNAs

are involved in many mammalian biological processes and also in the manifestation

of certain human diseases. As miRNAs play central role in the regulation of gene

expression, understanding miRNA-binding patterns is essential to gain an insight

of miRNA mediated gene regulation and also holds promise for therapeutic

applications. Computational prediction of miRNA binding sites on target mRNAs

facilitates experimental investigation of miRNA functions. This chapter provides

protocols for using the STarMir web server for improved predictions of miRNA

binding sites on a target mRNA. As an application module of the Sfold RNA

package, the current version of STarMir is an implementation of logistic

prediction models developed with high-throughput miRNA binding data from

cross-linking immunoprecipitation (CLIP) studies. The models incorporated

comprehensive thermodynamic, structural, and sequence features, and were found to

make improved predictions of both seed and seedless sites, in comparison to the

established algorithms (Liu et al., Nucleic Acids Res 41:e138, 2013). Their broad

applicability was indicated by their good performance in cross-species

validation. STarMir is freely available at

http://sfold.wadsworth.org/starmir.html .

DOI: 10.1007/978-1-4939-6433-8\_6

PMID: 27665594

378. Methods Mol Biol. 2016;1404:753-60. doi: 10.1007/978-1-4939-3389-1\_49.

MetaMHCpan, A Meta Approach for Pan-Specific MHC Peptide Binding Prediction.

Xu Y(1,)(2), Luo C(1,)(2), Mamitsuka H(3), Zhu S(4,)(5).

Author information:

(1)School of Computer Science, Fudan University, 220 Handan Road, Shanghai,

200433, China. (2)Shanghai Key Lab of Intelligent Information Processing, Fudan

University, Shanghai, 200433, China. (3)Bioinformatics Center, Institute for

Chemical Research, Kyoto University, Kyoto, 611-0011, Japan. (4)School of

Computer Science, Fudan University, 220 Handan Road, Shanghai, 200433, China.

zhusf@fudan.edu.cn. (5)Shanghai Key Lab of Intelligent Information Processing,

Fudan University, Shanghai, 200433, China. zhusf@fudan.edu.cn.

Recent computational approaches in bioinformatics can achieve high performance,

by which they can be a powerful support for performing real biological

experiments, making biologists pay more attention to bioinformatics than before.

In immunology, predicting peptides which can bind to MHC alleles is an important

task, being tackled by many computational approaches. However, this situation

causes a serious problem for immunologists to select the appropriate method to be

used in bioinformatics. To overcome this problem, we develop an ensemble

prediction-based Web server, which we call MetaMHCpan, consisting of two parts:

MetaMHCIpan and MetaMHCIIpan, for predicting peptides which can bind MHC-I and

MHC-II, respectively. MetaMHCIpan and MetaMHCIIpan use two (MHC2SKpan and LApan)

and four (TEPITOPEpan, MHC2SKpan, LApan, and MHC2MIL) existing predictors,

respectively. MetaMHCpan is available at

http://datamining-iip.fudan.edu.cn/MetaMHCpan/index.php/pages/view/info .

DOI: 10.1007/978-1-4939-3389-1\_49

PMID: 27076335 [Indexed for MEDLINE]

379. Methods Mol Biol. 2016;1399:207-33. doi: 10.1007/978-1-4939-3369-3\_13.

MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and

Function.

Keegan KP(1,)(2), Glass EM(3), Meyer F(4,)(5).

Author information:

(1)Argonne National Laboratory, 9700 South Cass Ave, Argonne, IL, 60439, USA.

kkeegan.argonne@gmail.com. (2)University of Chicago, Chicago, IL, USA.

kkeegan.argonne@gmail.com. (3)Argonne National Laboratory, 9700 South Cass Ave,

Argonne, IL, 60439, USA. marland@mcs.anl.gov. (4)Argonne National Laboratory,

9700 South Cass Ave, Argonne, IL, 60439, USA. folker@anl.gov. (5)University of

Chicago, Chicago, IL, USA. folker@anl.gov.

Approaches in molecular biology, particularly those that deal with

high-throughput sequencing of entire microbial communities (the field of

metagenomics), are rapidly advancing our understanding of the composition and

functional content of microbial communities involved in climate change,

environmental pollution, human health, biotechnology, etc. Metagenomics provides

researchers with the most complete picture of the taxonomic (i.e., what organisms

are there) and functional (i.e., what are those organisms doing) composition of

natively sampled microbial communities, making it possible to perform

investigations that include organisms that were previously intractable to

laboratory-controlled culturing; currently, these constitute the vast majority of

all microbes on the planet. All organisms contained in environmental samples are

sequenced in a culture-independent manner, most often with 16S ribosomal amplicon

methods to investigate the taxonomic or whole-genome shotgun-based methods to

investigate the functional content of sampled communities. Metagenomics allows

researchers to characterize the community composition and functional content of

microbial communities, but it cannot show which functional processes are active;

however, near parallel developments in transcriptomics promise a dramatic

increase in our knowledge in this area as well. Since 2008, MG-RAST (Meyer et

al., BMC Bioinformatics 9:386, 2008) has served as a public resource for

annotation and analysis of metagenomic sequence data, providing a repository that

currently houses more than 150,000 data sets (containing 60+ tera-base-pairs)

with more than 23,000 publically available. MG-RAST, or the metagenomics RAST

(rapid annotation using subsystems technology) server makes it possible for users

to upload raw metagenomic sequence data in (preferably) fastq or fasta format.

Assessments of sequence quality, annotation with respect to multiple reference

databases, are performed automatically with minimal input from the user (see

Subheading 4 at the end of this chapter for more details). Post-annotation

analysis and visualization are also possible, directly through the web interface,

or with tools like matR (metagenomic analysis tools for R, covered later in this

chapter) that utilize the MG-RAST API ( http://api.metagenomics.anl.gov/api.html

) to easily download data from any stage in the MG-RAST processing pipeline. Over

the years, MG-RAST has undergone substantial revisions to keep pace with the

dramatic growth in the number, size, and types of sequence data that accompany

constantly evolving developments in metagenomics and related -omic sciences

(e.g., metatranscriptomics).

DOI: 10.1007/978-1-4939-3369-3\_13

PMID: 26791506 [Indexed for MEDLINE]

380. Plant Cell Physiol. 2016 Jan;57(1):e2. doi: 10.1093/pcp/pcv189. Epub 2015 Dec 7.

The Vigna Genome Server, 'VigGS': A Genomic Knowledge Base of the Genus Vigna

Based on High-Quality, Annotated Genome Sequence of the Azuki Bean, Vigna

angularis (Willd.) Ohwi & Ohashi.

Sakai H(1), Naito K(2), Takahashi Y(2), Sato T(3), Yamamoto T(4), Muto I(4), Itoh

T(5), Tomooka N(2).

Author information:

(1)Agrogenomics Research Center, National Institute of Agrobiological Sciences,

2-1-2 Kannondai, Tsukuba, Ibaraki, 305-8602 Japan hirsakai@affrc.go.jp.

(2)Genetic Resources Center, National Institute of Agrobiological Sciences, 2-1-2

Kannondai, Tsukuba, Ibaraki, 305-8602 Japan. (3)Mizuho Information & Research

Institute, Inc., 2-3 Kanda-Nishikicho, Chiyoda-ku, Tokyo, 101-8443 Japan. (4)BITS

Co., Ltd., 5-201 Kandamatsunaga-cho, Chiyoda-ku, Tokyo, 101-0023 Japan.

(5)Agrogenomics Research Center, National Institute of Agrobiological Sciences,

2-1-2 Kannondai, Tsukuba, Ibaraki, 305-8602 Japan.

The genus Vigna includes legume crops such as cowpea, mungbean and azuki bean, as

well as >100 wild species. A number of the wild species are highly tolerant to

severe environmental conditions including high-salinity, acid or alkaline soil;

drought; flooding; and pests and diseases. These features of the genus Vigna make

it a good target for investigation of genetic diversity in adaptation to

stressful environments; however, a lack of genomic information has hindered such

research in this genus. Here, we present a genome database of the genus Vigna,

Vigna Genome Server ('VigGS', http://viggs.dna.affrc.go.jp), based on the

recently sequenced azuki bean genome, which incorporates annotated exon-intron

structures, along with evidence for transcripts and proteins, visualized in

GBrowse. VigGS also facilitates user construction of multiple alignments between

azuki bean genes and those of six related dicot species. In addition, the

database displays sequence polymorphisms between azuki bean and its wild

relatives and enables users to design primer sequences targeting any variant

site. VigGS offers a simple keyword search in addition to sequence similarity

searches using BLAST and BLAT. To incorporate up to date genomic information,

VigGS automatically receives newly deposited mRNA sequences of pre-set species

from the public database once a week. Users can refer to not only gene structures

mapped on the azuki bean genome on GBrowse but also relevant literature of the

genes. VigGS will contribute to genomic research into plant biotic and abiotic

stresses and to the future development of new stress-tolerant crops.

© The Author 2015. Published by Oxford University Press on behalf of Japanese

Society of Plant Physiologists. All rights reserved. For permissions, please

email: journals.permissions@oup.com.

DOI: 10.1093/pcp/pcv189

PMID: 26644460

381. RNA Biol. 2016;13(3):316-9. doi: 10.1080/15476286.2016.1141862. Epub 2016 Jan 29.

RiboGalaxy: A browser based platform for the alignment, analysis and

visualization of ribosome profiling data.

Michel AM(1), Mullan JP(1), Velayudhan V(1), O'Connor PB(1), Donohue CA(1),

Baranov PV(1).

Author information:

(1)a School of Biochemistry & Cell Biology, University College Cork , Cork ,

Ireland.

Ribosome profiling (ribo-seq) is a technique that uses high-throughput sequencing

to reveal the exact locations and densities of translating ribosomes at the

entire transcriptome level. The technique has become very popular since its

inception in 2009. Yet experimentalists who generate ribo-seq data often have to

rely on bioinformaticians to process and analyze their data. We present

RiboGalaxy ( http://ribogalaxy.ucc.ie ), a freely available Galaxy-based web

server for processing and analyzing ribosome profiling data with the

visualization functionality provided by GWIPS-viz ( http://gwips.ucc.ie ).

RiboGalaxy offers researchers a suite of tools specifically tailored for

processing ribo-seq and corresponding mRNA-seq data. Researchers can take

advantage of the published workflows which reduce the multi-step alignment

process to a minimum of inputs from the user. Users can then explore their own

aligned data as custom tracks in GWIPS-viz and compare their ribosome profiles to

existing ribo-seq tracks from published studies. In addition, users can assess

the quality of their ribo-seq data, determine the strength of the triplet

periodicity signal, generate meta-gene ribosome profiles as well as analyze the

relative impact of mRNA sequence features on local read density. RiboGalaxy is

accompanied by extensive documentation and tips for helping users. In addition we

provide a forum ( http://gwips.ucc.ie/Forum ) where we encourage users to post

their questions and feedback to improve the overall RiboGalaxy service.

DOI: 10.1080/15476286.2016.1141862

PMCID: PMC4829337

PMID: 26821742 [Indexed for MEDLINE]

382. PLoS One. 2015 Dec 29;10(12):e0145541. doi: 10.1371/journal.pone.0145541.

eCollection 2015.

iDPF-PseRAAAC: A Web-Server for Identifying the Defensin Peptide Family and

Subfamily Using Pseudo Reduced Amino Acid Alphabet Composition.

Zuo Y(1), Lv Y(1), Wei Z(1), Yang L(2), Li G(1), Fan G(3).

Author information:

(1)The Key Laboratory of Mammalian Reproductive Biology and Biotechnology of the

Ministry of Education, College of life sciences, Inner Mongolia University,

Hohhot, China. (2)College of Bioinformatics Science and Technology, Harbin

Medical University, Harbin, China. (3)Laboratory of Theoretical Biophysics,

School of Physical Science and Technology, Inner Mongolia University, Hohhot,

China.

Defensins as one of the most abundant classes of antimicrobial peptides are an

essential part of the innate immunity that has evolved in most living organisms

from lower organisms to humans. To identify specific defensins as interesting

antifungal leads, in this study, we constructed a more rigorous benchmark dataset

and the iDPF-PseRAAAC server was developed to predict the defensin family and

subfamily. Using reduced dipeptide compositions were used, the overall accuracy

of proposed method increased to 95.10% for the defensin family, and 98.39% for

the vertebrate subfamily, which is higher than the accuracy from other methods.

The jackknife test shows that more than 4% improvement was obtained comparing

with the previous method. A free online server was further established for the

convenience of most experimental scientists at

http://wlxy.imu.edu.cn/college/biostation/fuwu/iDPF-PseRAAAC/index.asp. A

friendly guide is provided to describe how to use the web server. We anticipate

that iDPF-PseRAAAC may become a useful high-throughput tool for both basic

research and drug design.

DOI: 10.1371/journal.pone.0145541

PMCID: PMC4694767

PMID: 26713618 [Indexed for MEDLINE]

383. Anal Biochem. 2015 Dec 15;491:18-22. doi: 10.1016/j.ab.2015.08.028. Epub 2015 Sep

6.

Prediction of protein disorder on amino acid substitutions.

Anoosha P(1), Sakthivel R(1), Gromiha MM(2).

Author information:

(1)Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences,

Indian Institute of Technology Madras, Chennai 600036, Tamilnadu, India.

(2)Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences,

Indian Institute of Technology Madras, Chennai 600036, Tamilnadu, India.

Electronic address: gromiha@iitm.ac.in.

Intrinsically disordered regions of proteins are known to have many functional

roles in cell signaling and regulatory pathways. The altered expression of these

proteins due to mutations is associated with various diseases. Currently, most of

the available methods focus on predicting the disordered proteins or the

disordered regions in a protein. On the other hand, methods developed for

predicting protein disorder on mutation showed a poor performance with a maximum

accuracy of 70%. Hence, in this work, we have developed a novel method to

classify the disorder-related amino acid substitutions using amino acid

properties, substitution matrices, and the effect of neighboring residues that

showed an accuracy of 90.0% with a sensitivity and specificity of 94.9 and 80.6%,

respectively, in 10-fold cross-validation. The method was evaluated with a test

set of 20% data using 10 iterations, which showed an average accuracy of 88.9%.

Furthermore, we systematically analyzed the features responsible for the better

performance of our method and observed that neighboring residues play an

important role in defining the disorder of a given residue in a protein sequence.

We have developed a prediction server to identify disorder-related mutations, and

it is available at http://www.iitm.ac.in/bioinfo/DIM\_Pred/.

Copyright © 2015 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2015.08.028

PMID: 26348538 [Indexed for MEDLINE]

384. Bioinformatics. 2015 Dec 15;31(24):4038-40. doi: 10.1093/bioinformatics/btv503.

Epub 2015 Aug 30.

RNASeqMetaDB: a database and web server for navigating metadata of publicly

available mouse RNA-Seq datasets.

Guo Z(1), Tzvetkova B(2), Bassik JM(3), Bodziak T(3), Wojnar BM(3), Qiao W(4),

Obaida MA(4), Nelson SB(2), Hu BH(3), Yu P(1).

Author information:

(1)Department of Electrical and Computer Engineering & TEES-AgriLife Center for

Bioinformatics and Genomic Systems Engineering, Texas A&M University, College

Station, TX 77843, USA. (2)Department of Biology & Center for Behavioral

Genomics, Brandeis University, Waltham, MA 02454, USA and. (3)Department of

Communicative Disorders and Sciences & Center for Hearing and Deafness,

University at Buffalo, Buffalo, NY 14214, USA. (4)Department of Electrical and

Computer Engineering &

Gene targeting is a protocol for introducing a mutation to a specific gene in an

organism. Because of the importance of in vivo assessment of gene function and

modeling of human diseases, this technique has been widely adopted to generate a

large number of mutant mouse models. Due to the recent breakthroughs in

high-throughput sequencing technologies, RNA-Seq experiments have been performed

on many of these mouse models, leading to hundreds of publicly available

datasets. To facilitate the reuse of these datasets, we collected the associated

metadata and organized them in a database called RNASeqMetaDB. The metadata were

manually curated to ensure annotation consistency. We developed a web server to

allow easy database navigation and data querying. Users can search the database

using multiple parameters like genes, diseases, tissue types, keywords and

associated publications in order to find datasets that match their interests.

Summary statistics of the metadata are also presented on the web server showing

interesting global patterns of RNA-Seq studies.AVAILABILITY AND IMPLEMENTATION:

Freely available on the web at http://rnaseqmetadb.ece.tamu.edu.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv503

PMCID: PMC4692969

PMID: 26323714 [Indexed for MEDLINE]

385. J Comput Chem. 2015 Dec 15;36(32):2381-93. doi: 10.1002/jcc.24218. Epub 2015 Oct

20.

Statistical investigation of surface bound ions and further development of BION

server to include pH and salt dependence.

Petukh M(1), Zhang M(1), Alexov E(1).

Author information:

(1)Computational Biophysics and Bioinformatics, Department of Physics and

Astronomy, Clemson University, Clemson, SC, 29634.

Ions are engaged in multiple biological processes in cells. By binding to the

macromolecules or being mobile in the solvent, they maintain the integrity of the

structure of macromolecules; participate in their enzymatic activity; or screen

electrostatic interactions. While experimental methods are not always able to

assign the exact location of ions, computational methods are in demand. Although

the majority of computational methods are successful in predicting the position

of ions buried inside macromolecules, they are less effective in deciphering

positions of surface bound ions. Here, we propose the new BION algorithm

(http://compbio.clemson.edu/bion\_server\_ph/) that predicts the location of the

surface bound ions. It is more efficient and accurate compared to the previous

version since it uses more advanced clustering algorithm in combination with

pairing rules. In addition, the BION webserver allows specifying the pH and the

salt concentration in predicting ions positions.

© 2015 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24218

PMID: 26484964 [Indexed for MEDLINE]

386. Sci Rep. 2015 Dec 9;5:16964. doi: 10.1038/srep16964.

Predicting cancerlectins by the optimal g-gap dipeptides.

Lin H(1), Liu WX(1), He J(1), Liu XH(2), Ding H(1), Chen W(1,)(3).

Author information:

(1)Key Laboratory for Neuro-Information of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, Center for Information in

Biomedicine, University of Electronic Science and Technology of China, Chengdu

610054, China. (2)School of Linguistics and Literature, University of Electronic

Science and Technology of China, Chengdu 610054, China. (3)Department of Physics,

School of Sciences, and Center for Genomics and Computational Biology, North

China University of Science and Technology, Tangshan 063000, China.

The cancerlectin plays a key role in the process of tumor cell differentiation.

Thus, to fully understand the function of cancerlectin is significant because it

sheds light on the future direction for the cancer therapy. However, the

traditional wet-experimental methods were money- and time-consuming. It is highly

desirable to develop an effective and efficient computational tool to identify

cancerlectins. In this study, we developed a sequence-based method to

discriminate between cancerlectins and non-cancerlectins. The analysis of

variance (ANOVA) was used to choose the optimal feature set derived from the

g-gap dipeptide composition. The jackknife cross-validated results showed that

the proposed method achieved the accuracy of 75.19%, which is superior to other

published methods. For the convenience of other researchers, an online web-server

CaLecPred was established and can be freely accessed from the website

http://lin.uestc.edu.cn/server/CalecPred. We believe that the CaLecPred is a

powerful tool to study cancerlectins and to guide the related experimental

validations.

DOI: 10.1038/srep16964

PMCID: PMC4673586

PMID: 26648527 [Indexed for MEDLINE]

387. Sci Rep. 2015 Dec 4;5:17573. doi: 10.1038/srep17573.

Improving Protein Fold Recognition by Deep Learning Networks.

Jo T(1,)(2), Hou J(1), Eickholt J(3), Cheng J(1).

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, MO 65211,

USA. (2)Department of Biological Chemistry, University of Michigan, Ann Arbor,

MI, 48109, USA. (3)Department of Computer Science, Central Michigan University,

Mount Pleasant, MI 48859, USA.

For accurate recognition of protein folds, a deep learning network method

(DN-Fold) was developed to predict if a given query-template protein pair belongs

to the same structural fold. The input used stemmed from the protein sequence and

structural features extracted from the protein pair. We evaluated the performance

of DN-Fold along with 18 different methods on Lindahl's benchmark dataset and on

a large benchmark set extracted from SCOP 1.75 consisting of about one million

protein pairs, at three different levels of fold recognition (i.e., protein

family, superfamily, and fold) depending on the evolutionary distance between

protein sequences. The correct recognition rate of ensembled DN-Fold for Top 1

predictions is 84.5%, 61.5%, and 33.6% and for Top 5 is 91.2%, 76.5%, and 60.7%

at family, superfamily, and fold levels, respectively. We also evaluated the

performance of single DN-Fold (DN-FoldS), which showed the comparable results at

the level of family and superfamily, compared to ensemble DN-Fold. Finally, we

extended the binary classification problem of fold recognition to real-value

regression task, which also show a promising performance. DN-Fold is freely

available through a web server at http://iris.rnet.missouri.edu/dnfold.

DOI: 10.1038/srep17573

PMCID: PMC4669437

PMID: 26634993 [Indexed for MEDLINE]

388. IEEE/ACM Trans Comput Biol Bioinform. 2015 Dec 3. [Epub ahead of print]

PyMut: a web tool for overlapping gene loss-of-function mutation design.

Liu K, Hou S, Dai J, Sun Z.

Loss-of-function study is an effective approach to research gene functions.

However, currently most of such studies have ignored an important problem (in

this paper, we call it "off-target" problem), that is, if the target gene is an

overlapping gene (A gene whose expressible nucleotides overlaps with that of

another one), loss-of-function muta-tion by deleting the complete open reading

frame (ORF) may also cause the gene it overlaps lose function, resulting a

phenotype which may be rather different from that of single gene deletion.

Therefore, when doing such studies, the loss-of-function mutations should be

carefully designed to guarantee only the function of the target gene will be

abolished. In this paper, we present PyMut, an easy-to-use web tool for

biologists to design such mutations. To the best of our knowledge, PyMut is the

first tool that aims to solve the "off-target" problem regarding the overlapping

genes. Our web server is freely available at

http://www.bioinfo.tsinghua.edu.cn/~liuke/PyMut/index.html.

DOI: 10.1109/TCBB.2015.2505290

PMID: 26661787

389. Anal Biochem. 2015 Dec 1;490:26-33. doi: 10.1016/j.ab.2015.08.021. Epub 2015 Aug

24.

iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide

composition.

Chen W(1), Feng P(2), Ding H(2), Lin H(3), Chou KC(4).

Author information:

(1)Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, North China University of Science and Technology, Tangshan

063009, China; Gordon Life Science Institute, Belmont, MA 02478, USA. Electronic

address: chenweiimu@gmail.com. (2)School of Public Health, North China University

of Science and Technology, Tangshan 063000, China. (3)Gordon Life Science

Institute, Belmont, MA 02478, USA; Key Laboratory for Neuro-Information of

Ministry of Education, Center of Bioinformatics and Center for Information in

Biomedicine, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu 610054, China. Electronic address:

hlin@uestc.edu.cn. (4)Gordon Life Science Institute, Belmont, MA 02478, USA;

Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz

University, Jeddah 21589, Saudi Arabia. Electronic address:

kcchou@gordonlifescience.org.

Occurring at adenine (A) with the consensus motif GAC, N(6)-methyladenosine

(m(6)A) is one of the most abundant modifications in RNA, which plays very

important roles in many biological processes. The nonuniform distribution of

m(6)A sites across the genome implies that, for better understanding the

regulatory mechanism of m(6)A, it is indispensable to characterize its sites in a

genome-wide scope. Although a series of experimental technologies have been

developed in this regard, they are both time-consuming and expensive. With the

avalanche of RNA sequences generated in the postgenomic age, it is highly desired

to develop computational methods to timely identify their m(6)A sites. In view of

this, a predictor called "iRNA-Methyl" is proposed by formulating RNA sequences

with the "pseudo dinucleotide composition" into which three RNA physiochemical

properties were incorporated. Rigorous cross-validation tests have indicated that

iRNA-Methyl holds very high potential to become a useful tool for genome

analysis. For the convenience of most experimental scientists, a web-server for

iRNA-Methyl has been established at http://lin.uestc.edu.cn/server/iRNA-Methyl by

which users can easily get their desired results without needing to go through

the mathematical details.

Copyright © 2015 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2015.08.021

PMID: 26314792 [Indexed for MEDLINE]

390. Bioinformatics. 2015 Dec 1;31(23):3748-50. doi: 10.1093/bioinformatics/btv439.

Epub 2015 Aug 10.

SuccFind: a novel succinylation sites online prediction tool via enhanced

characteristic strategy.

Xu HD(1), Shi SP(2), Wen PP(1), Qiu JD(3).

Author information:

(1)Department of Chemistry. (2)Department of Mathematics, Nanchang University,

Nanchang 330031, China and. (3)Department of Chemistry, Department of Chemical

Engineering, Pingxiang University, Pingxiang 337055, China.

Lysine succinylation orchestrates a variety of biological processes. Annotation

of succinylation in proteomes is the first-crucial step to decipher physiological

roles of succinylation implicated in the pathological processes. In this work, we

developed a novel succinylation site online prediction tool, called SuccFind,

which is constructed to predict the lysine succinylation sites based on two major

categories of characteristics: sequence-derived features and evolutionary-derived

information of sequence and via an enhanced feature strategy for further

optimizations. The assessment results obtained from cross-validation suggest that

SuccFind can provide more instructive guidance for further experimental

investigation of protein succinylation.AVAILABILITY AND IMPLEMENTATION: A

user-friendly server is freely available on the web at:

http://bioinfo.ncu.edu.cn/SuccFind.aspx.

CONTACT: jdqiu@ncu.edu.cn.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv439

PMID: 26261224 [Indexed for MEDLINE]

391. Bioinformatics. 2015 Dec 1;31(23):3850-2. doi: 10.1093/bioinformatics/btv441.

Epub 2015 Jul 30.

HHalign-Kbest: exploring sub-optimal alignments for remote homology comparative

modeling.

Yu J(1), Picord G(2), Tuffery P(2), Guerois R(1).

Author information:

(1)Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University

Paris-Saclay, CEA-Saclay, 91191 Gif-sur-Yvette. (2)INSERM U973, MTi, F-75205

Paris, Université Paris Diderot, Sorbonne Paris Cité F-75205 Paris and Ressource

Parisienne en Bioinformatique Structurale, F-75205 Paris, France.

MOTIVATION: The HHsearch algorithm, implementing a hidden Markov model (HMM)-HMM

alignment method, has shown excellent alignment performance in the so-called

twilight zone (target-template sequence identity with ∼20%). However, an optimal

alignment by HHsearch may contain small to large errors, leading to poor

structure prediction if these errors are located in important structural

elements.

RESULTS: HHalign-Kbest server runs a full pipeline, from the generation of

suboptimal HMM-HMM alignments to the evaluation of the best structural models. In

the HHsearch framework, it implements a novel algorithm capable of generating

k-best HMM-HMM suboptimal alignments rather than only the optimal one. For large

proteins, a directed acyclic graph-based implementation reduces drastically the

memory usage. Improved alignments were systematically generated among the top k

suboptimal alignments. To recognize them, corresponding structural models were

systematically generated and evaluated with Qmean score. The method was

benchmarked over 420 targets from the SCOP30 database. In the range of HHsearch

probability of 20-99%, average quality of the models (TM-score) raised by

4.1-16.3% and 8.0-21.0% considering the top 1 and top 10 best models,

respectively.

AVAILABILITY AND IMPLEMENTATION:

http://bioserv.rpbs.univ-paris-diderot.fr/services/HHalign-Kbest/ (source code

and server).

CONTACT: guerois@cea.fr.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv441

PMID: 26231431 [Indexed for MEDLINE]

392. BMC Bioinformatics. 2015 Dec 1;16:400. doi: 10.1186/s12859-015-0827-2.

PC-TraFF: identification of potentially collaborating transcription factors using

pointwise mutual information.

Meckbach C(1), Tacke R(2), Hua X(3), Waack S(4), Wingender E(5), Gültas M(6).

Author information:

(1)Institute of Bioinformatics, University of Göttingen, Goldschmidtstr. 1,

Göttingen, 37077, Germany. cornelia.meckbach@bioinf.med.uni-goettingen.de.

(2)Institute of Bioinformatics, University of Göttingen, Goldschmidtstr. 1,

Göttingen, 37077, Germany. rebecca.tacke@stud.uni-goettingen.de. (3)Institute of

Bioinformatics, University of Göttingen, Goldschmidtstr. 1, Göttingen, 37077,

Germany. xu.hua@bioinf.med.uni-goettingen.de. (4)Institute of Computer Science,

University of Göttingen, Goldschmidtstr. 7, Göttingen, 37077, Germany.

waack@cs.uni-goettingen.de. (5)Institute of Bioinformatics, University of

Göttingen, Goldschmidtstr. 1, Göttingen, 37077, Germany.

edgar.wingender@bioinf.med.uni-goettingen.de. (6)Institute of Bioinformatics,

University of Göttingen, Goldschmidtstr. 1, Göttingen, 37077, Germany.

mehmet.gueltas@bioinf.med.uni-goettingen.de.

BACKGROUND: Transcription factors (TFs) are important regulatory proteins that

govern transcriptional regulation. Today, it is known that in higher organisms

different TFs have to cooperate rather than acting individually in order to

control complex genetic programs. The identification of these interactions is an

important challenge for understanding the molecular mechanisms of regulating

biological processes. In this study, we present a new method based on pointwise

mutual information, PC-TraFF, which considers the genome as a document, the

sequences as sentences, and TF binding sites (TFBSs) as words to identify

interacting TFs in a set of sequences.

RESULTS: To demonstrate the effectiveness of PC-TraFF, we performed a genome-wide

analysis and a breast cancer-associated sequence set analysis for protein coding

and miRNA genes. Our results show that in any of these sequence sets, PC-TraFF is

able to identify important interacting TF pairs, for most of which we found

support by previously published experimental results. Further, we made a pairwise

comparison between PC-TraFF and three conventional methods. The outcome of this

comparison study strongly suggests that all these methods focus on different

important aspects of interaction between TFs and thus the pairwise overlap

between any of them is only marginal.

CONCLUSIONS: In this study, adopting the idea from the field of linguistics in

the field of bioinformatics, we develop a new information theoretic method,

PC-TraFF, for the identification of potentially collaborating transcription

factors based on the idiosyncrasy of their binding site distributions on the

genome. The results of our study show that PC-TraFF can succesfully identify

known interacting TF pairs and thus its currently biologically uncorfirmed

predictions could provide new hypotheses for further experimental validation.

Additionally, the comparison of the results of PC-TraFF with the results of

previous methods demonstrates that different methods with their specific scopes

can perfectly supplement each other. Overall, our analyses indicate that PC-TraFF

is a time-efficient method where its algorithm has a tractable computational time

and memory consumption. The PC-TraFF server is freely accessible at

http://pctraff.bioinf.med.uni-goettingen.de/.

DOI: 10.1186/s12859-015-0827-2

PMCID: PMC4667426

PMID: 26627005 [Indexed for MEDLINE]

393. Eur J Med Genet. 2015 Dec;58(12):715-8. doi: 10.1016/j.ejmg.2015.10.007. Epub

2015 Oct 24.

Penetrance of pathogenic mutations in haploinsufficient genes for intellectual

disability and related disorders.

Ropers HH(1), Wienker T(2).

Author information:

(1)Institute for Human Genetics, University Medicine, Langenbeckstrasse 1,

Building 601, 55131 Mainz, Germany; Max Planck Institute for Molecular Genetics,

Ihnestrasse 73, 14195 Berlin, Germany. Electronic address: ropers@molgen.mpg.de.

(2)Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin,

Germany.

De novo loss of function (LOF) mutations in the ASXL3 gene cause

Bainbridge-Ropers syndrome, a severe form of intellectual disability (ID) and

developmental delay, but there is evidence that they also occur in healthy

individuals. This has prompted us to look for non-pathogenic LOF variants in

other ID genes. Heterozygous LOF mutations in ASXL1, a paralog of ASXL3, are

known to cause Bohring-Opitz syndrome (BOS), and benign LOF mutations in this

gene have not been published to date. Therefore, we were surprised to find 56

ASXL1 LOF variants in the ExAC database (http://exac.broadinstitute.org),

comprising exomes from 60,706 individuals who had been selected to exclude severe

genetic childhood disorders. 4 of these variants have been described as

disease-causing in patients with BOS, which rules out the possibility that

pathogenic and clinically neutral LOF variants in this gene are functionally

distinct. Apparently benign LOF variants were also detected in several other

genes for ID and related disorders, including CDH15, KATNAL2, DEPDC5, ARID1B and

AUTS2, both in the ExAC database and in the 6,500 exomes of the Exome Variant

Server (http://evs.gs.washington.edu/EVS/). These observations argue for low

penetrance of LOF mutations in ASXL1 and other genes for ID and related

disorders, which could have far-reaching implications for genetic counseling and

research.

Copyright © 2015 Elsevier Masson SAS. All rights reserved.

DOI: 10.1016/j.ejmg.2015.10.007

PMID: 26506440 [Indexed for MEDLINE]

394. J Membr Biol. 2015 Dec;248(6):1033-41. doi: 10.1007/s00232-015-9815-8. Epub 2015

Jun 16.

iCataly-PseAAC: Identification of Enzymes Catalytic Sites Using Sequence

Evolution Information with Grey Model GM (2,1).

Xiao X(1,)(2,)(3), Hui MJ(4), Liu Z(5), Qiu WR(6).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333046,

China. jdzxiaoxuan@163.com. (2)Information School, ZheJiang Textile & Fashion

College, NingBo, 315211, China. jdzxiaoxuan@163.com. (3)Gordon Life Science

Institute, 53 South Cottage Road, Belmont, MA, 02478, USA. jdzxiaoxuan@163.com.

(4)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333046,

China. huimengjuan@163.com. (5)Computer Department, Jing-De-Zhen Ceramic

Institute, Jing-De-Zhen, 333046, China. liuzi189836@163.com. (6)Computer

Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333046, China.

qiuone@163.com.

Enzymes play pivotal roles in most of the biological reaction. The catalytic

residues of an enzyme are defined as the amino acids which are directly involved

in chemical catalysis; the knowledge of these residues is important for

understanding enzyme function. Given an enzyme, which residues are the catalytic

sites, and which residues are not? This is the first important problem for

in-depth understanding the catalytic mechanism and drug development. With the

explosive of protein sequences generated during the post-genomic era, it is

highly desirable for both basic research and drug design to develop fast and

reliable method for identifying the catalytic sites of enzymes according to their

sequences. To address this problem, we proposed a new predictor, called

iCataly-PseAAC. In the prediction system, the peptide sample was formulated with

sequence evolution information via grey system model GM(2,1). It was observed by

the rigorous jackknife test and independent dataset test that iCataly-PseAAC was

superior to exist predictions though its only use sequence information. As a

user-friendly web server, iCataly-PseAAC is freely accessible at

http://www.jci-bioinfo.cn/iCataly-PseAAC. A step-by-step guide has been provided

on how to use the web server to get the desired results for the convenience of

most experimental scientists.

DOI: 10.1007/s00232-015-9815-8

PMID: 26077845 [Indexed for MEDLINE]

395. J Membr Biol. 2015 Dec;248(6):1005-14. doi: 10.1007/s00232-015-9811-z. Epub 2015

Jun 10.

TargetFreeze: Identifying Antifreeze Proteins via a Combination of Weights using

Sequence Evolutionary Information and Pseudo Amino Acid Composition.

He X(1), Han K(1), Hu J(1), Yan H(1), Yang JY(1), Shen HB(2), Yu DJ(3,)(4).

Author information:

(1)School of Computer Science and Engineering, Nanjing University of Science and

Technology, Xiaolingwei 200, Nanjing, 210094, China. (2)Institute of Image

Processing and Pattern Recognition, Shanghai Jiao Tong University, Dongchuan Road

800, Shanghai, 200240, China. (3)School of Computer Science and Engineering,

Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, 210094,

China. njyudj@njust.edu.cn. (4)Changshu Institute, Nanjing University of Science

and Technology, Changshu, 215513, China. njyudj@njust.edu.cn.

Antifreeze proteins (AFPs) are indispensable for living organisms to survive in

an extremely cold environment and have a variety of potential biotechnological

applications. The accurate prediction of antifreeze proteins has become an

important issue and is urgently needed. Although considerable progress has been

made, AFP prediction is still a challenging problem due to the diversity of

species. In this study, we proposed a new sequence-based AFP predictor, called

TargetFreeze. TargetFreeze utilizes an enhanced feature representation method

that weightedly combines multiple protein features and takes the powerful support

vector machine as the prediction engine. Computer experiments on benchmark

datasets demonstrate the superiority of the proposed TargetFreeze over most

recently released AFP predictors. We also implemented a user-friendly web server,

which is openly accessible for academic use and is available at

http://csbio.njust.edu.cn/bioinf/TargetFreeze. TargetFreeze supplements existing

AFP predictors and will have potential applications in AFP-related biotechnology

fields.

DOI: 10.1007/s00232-015-9811-z

PMID: 26058944 [Indexed for MEDLINE]

396. Rev Sci Instrum. 2015 Dec;86(12):125003. doi: 10.1063/1.4937617.

Development of intelligent instruments with embedded HTTP servers for control and

data acquisition in a cryogenic setup--The hardware, firmware, and software

implementation.

Antony J(1), Mathuria DS(1), Datta TS(1), Maity T(2).

Author information:

(1)Inter University Accelerator Centre, Aruna Asaf Ali Marg, New Delhi 110067,

India. (2)Department of MME, Indian School of Mines (ISM), Dhanbad 826004, India.

The power of Ethernet for control and automation technology is being largely

understood by the automation industry in recent times. Ethernet with HTTP

(Hypertext Transfer Protocol) is one of the most widely accepted communication

standards today. Ethernet is best known for being able to control through

internet from anywhere in the globe. The Ethernet interface with built-in on-chip

embedded servers ensures global connections for crate-less model of control and

data acquisition systems which have several advantages over traditional

crate-based control architectures for slow applications. This architecture will

completely eliminate the use of any extra PLC (Programmable Logic Controller) or

similar control hardware in any automation network as the control functions are

firmware coded inside intelligent meters itself. Here, we describe the

indigenously built project of a cryogenic control system built for linear

accelerator at Inter University Accelerator Centre, known as "CADS," which stands

for "Complete Automation of Distribution System." CADS deals with complete

hardware, firmware, and software implementation of the automated linac cryogenic

distribution system using many Ethernet based embedded cryogenic instruments

developed in-house. Each instrument works as an intelligent meter called

device-server which has the control functions and control loops built inside the

firmware itself. Dedicated meters with built-in servers were designed out of ARM

(Acorn RISC (Reduced Instruction Set Computer) Machine) and ATMEL processors and

COTS (Commercially Off-the-Shelf) SMD (Surface Mount Devices) components, with

analog sensor front-end and a digital back-end web server implementing remote

procedure call over HTTP for digital control and readout functions. At present,

24 instruments which run 58 embedded servers inside, each specific to a

particular type of sensor-actuator combination for closed loop operations, are

now deployed and distributed across control LAN (Local Area Network). A group of

six categories of such instruments have been identified for all cryogenic

applications required for linac operation which were designed to build this

medium-scale cryogenic automation setup. These devices have special features like

remote rebooters, daughter boards for PIDs (Proportional Integral Derivative),

etc., to operate them remotely in radiation areas and also have emergency

switches by which each device can be taken to emergency mode temporarily.

Finally, all the data are monitored, logged, controlled, and analyzed online at a

central control room which has a user-friendly control interface developed using

LabVIEW(®). This paper discusses the overall hardware, firmware, software design,

and implementation for the cryogenics setup.

DOI: 10.1063/1.4937617

PMID: 26724062

397. Sci Rep. 2015 Nov 26;5:17155. doi: 10.1038/srep17155.

Inferring the hosts of coronavirus using dual statistical models based on

nucleotide composition.

Tang Q(1,)(2), Song Y(1,)(2), Shi M(1), Cheng Y(1), Zhang W(1), Xia XQ(1).

Author information:

(1)Center for Molecular and Cellular Biology of Aquatic Organisms, Institute of

Hydrobiology, the Chinese Academy of Sciences, Wuhan 430072, China. (2)University

of Chinese Academy of Sciences, Beijing, China.

Many coronaviruses are capable of interspecies transmission. Some of them have

caused worldwide panic as emerging human pathogens in recent years, e.g., severe

acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory

syndrome coronavirus (MERS-CoV). In order to assess their threat to humans, we

explored to infer the potential hosts of coronaviruses using a dual-model

approach based on nineteen parameters computed from spike genes of coronaviruses.

Both the support vector machine (SVM) model and the Mahalanobis distance (MD)

discriminant model achieved high accuracies in leave-one-out cross-validation of

training data consisting of 730 representative coronaviruses (99.86% and 98.08%

respectively). Predictions on 47 additional coronaviruses precisely conformed to

conclusions or speculations by other researchers. Our approach is implemented as

a web server that can be accessed at http://bioinfo.ihb.ac.cn/seq2hosts.

DOI: 10.1038/srep17155

PMCID: PMC4660426

PMID: 26607834 [Indexed for MEDLINE]

398. J Chem Inf Model. 2015 Nov 23;55(11):2308-14. doi: 10.1021/acs.jcim.5b00534. Epub

2015 Nov 9.

ProBiS-CHARMMing: Web Interface for Prediction and Optimization of Ligands in

Protein Binding Sites.

Konc J(1,)(2), Miller BT(3), Štular T(2), Lešnik S(1), Woodcock HL(4), Brooks

BR(3), Janežič D(2).

Author information:

(1)Laboratory for Molecular Modeling, National Institute of Chemistry ,

Hajdrihova 19, SI-1000, Ljubljana, Slovenia. (2)Faculty of Mathematics, Natural

Sciences and Information Technologies, University of Primorska , Glagoljaška 8,

SI-6000, Koper, Slovenia. (3)Laboratory of Computational Biology, Biochemistry

and Biophysics Center, National Heart, Lung and Blood Institute, National

Institutes of Health , Bethesda, Maryland 20892, United States. (4)Department of

Chemistry, University of South Florida , 4202 East Fowler Ave., Tampa, Florida

33620, United States.

Proteins often exist only as apo structures (unligated) in the Protein Data Bank,

with their corresponding holo structures (with ligands) unavailable. However,

apoproteins may not represent the amino-acid residue arrangement upon ligand

binding well, which is especially problematic for molecular docking. We developed

the ProBiS-CHARMMing web interface by connecting the ProBiS (

http://probis.cmm.ki.si ) and CHARMMing ( http://www.charmming.org ) web servers

into one functional unit that enables prediction of protein-ligand complexes and

allows for their geometry optimization and interaction energy calculation. The

ProBiS web server predicts ligands (small compounds, proteins, nucleic acids, and

single-atom ligands) that may bind to a query protein. This is achieved by

comparing its surface structure against a nonredundant database of protein

structures and finding those that have binding sites similar to that of the query

protein. Existing ligands found in the similar binding sites are then transposed

to the query according to predictions from ProBiS. The CHARMMing web server

enables, among other things, minimization and potential energy calculation for a

wide variety of biomolecular systems, and it is used here to optimize the

geometry of the predicted protein-ligand complex structures using the CHARMM

force field and to calculate their interaction energies with the corresponding

query proteins. We show how ProBiS-CHARMMing can be used to predict ligands and

their poses for a particular binding site, and minimize the predicted

protein-ligand complexes to obtain representations of holoproteins. The

ProBiS-CHARMMing web interface is freely available for academic users at

http://probis.nih.gov.

DOI: 10.1021/acs.jcim.5b00534

PMID: 26509288 [Indexed for MEDLINE]

399. J Theor Biol. 2015 Nov 21;385:153-9. doi: 10.1016/j.jtbi.2015.08.025. Epub 2015

Sep 9.

Identification of microRNA precursor with the degenerate K-tuple or Kmer

strategy.

Liu B(1), Fang L(2), Wang S(3), Wang X(4), Li H(5), Chou KC(6).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong, China; Key Laboratory of Network

Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, Guangdong, China; Gordon Life Science Institute,

Boston, MA 0478, USA. Electronic address: bliu@insun.hit.edu.cn. (2)School of

Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate

School, Shenzhen, Guangdong, China. Electronic address: dragoncloudest@gmail.com.

(3)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong, China. Electronic address:

wangshanyiwsy@gmail.com. (4)School of Computer Science and Technology, Harbin

Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China; Key

Laboratory of Network Oriented Intelligent Computation, Harbin Institute of

Technology Shenzhen Graduate School, Shenzhen, Guangdong, China. Electronic

address: wangxl@insun.hit.edu.cn. (5)Wendeng Marine Environmental Monitoring

Station, Station Oceanic Administration Wendeng, Weihai, Shandong, China.

Electronic address: lht760@126.com. (6)Gordon Life Science Institute, Boston, MA

0478, USA; Center of Excellence in Genomic Medicine Research (CEGMR), King

Abdulaziz University, Jeddah 21589, Saudi Arabia. Electronic address:

kcchou@gordonlifescience.org.

The microRNA (miRNA), a small non-coding RNA molecule, plays an important role in

transcriptional and post-transcriptional regulation of gene expression. Its

abnormal expression, however, has been observed in many cancers and other disease

states, implying that the miRNA molecules are also deeply involved in these

diseases, particularly in carcinogenesis. Therefore, it is important for both

basic research and miRNA-based therapy to discriminate the real pre-miRNAs from

the false ones (such as hairpin sequences with similar stem-loops). Most existing

methods in this regard were based on the strategy in which RNA samples were

formulated by a vector formed by their Kmer components. But the length of Kmers

must be very short; otherwise, the vector's dimension would be extremely large,

leading to the "high-dimension disaster" or overfitting problem. Inspired by the

concept of "degenerate energy levels" in quantum mechanics, we introduced the

"degenerate Kmer" (deKmer) to represent RNA samples. By doing so, not only we can

accommodate long-range coupling effects but also we can avoid the high-dimension

problem. Rigorous jackknife tests and cross-species experiments indicated that

our approach is very promising. It has not escaped our notice that the deKmer

approach can also be applied to many other areas of computational biology. A

user-friendly web-server for the new predictor has been established at

http://bioinformatics.hitsz.edu.cn/miRNA-deKmer/, by which users can easily get

their desired results.

Copyright © 2015 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2015.08.025

PMID: 26362104 [Indexed for MEDLINE]

400. Source Code Biol Med. 2015 Nov 19;10:14. doi: 10.1186/s13029-015-0042-6.

eCollection 2015.

PyPedia: using the wiki paradigm as crowd sourcing environment for bioinformatics

protocols.

Kanterakis A(1), Kuiper J(2), Potamias G(3), Swertz MA(2).

Author information:

(1)University of Groningen, University Medical Center Groningen, Genomics

Coordination Center, Postbus 30 001, Groningen, 9700 RB The Netherlands ;

Institute of Computer Science, Foundation for Research and Technology Hellas

(FORTH), Nikolaou Plastira 100, Heraklion, 71110 Greece. (2)University of

Groningen, University Medical Center Groningen, Genomics Coordination Center,

Postbus 30 001, Groningen, 9700 RB The Netherlands. (3)Institute of Computer

Science, Foundation for Research and Technology Hellas (FORTH), Nikolaou Plastira

100, Heraklion, 71110 Greece.

BACKGROUND: Today researchers can choose from many bioinformatics protocols for

all types of life sciences research, computational environments and coding

languages. Although the majority of these are open source, few of them possess

all virtues to maximize reuse and promote reproducible science. Wikipedia has

proven a great tool to disseminate information and enhance collaboration between

users with varying expertise and background to author qualitative content via

crowdsourcing. However, it remains an open question whether the wiki paradigm can

be applied to bioinformatics protocols.

RESULTS: We piloted PyPedia, a wiki where each article is both implementation and

documentation of a bioinformatics computational protocol in the python language.

Hyperlinks within the wiki can be used to compose complex workflows and induce

reuse. A RESTful API enables code execution outside the wiki. Initial content of

PyPedia contains articles for population statistics, bioinformatics format

conversions and genotype imputation. Use of the easy to learn wiki syntax

effectively lowers the barriers to bring expert programmers and less computer

savvy researchers on the same page.

CONCLUSIONS: PyPedia demonstrates how wiki can provide a collaborative

development, sharing and even execution environment for biologists and

bioinformaticians that complement existing resources, useful for local and

multi-center research teams.

AVAILABILITY: PyPedia is available online at: http://www.pypedia.com. The source

code and installation instructions are available at:

https://github.com/kantale/PyPedia\_server. The PyPedia python library is

available at: https://github.com/kantale/pypedia. PyPedia is open-source,

available under the BSD 2-Clause License.

DOI: 10.1186/s13029-015-0042-6

PMCID: PMC4652372

PMID: 26587054

401. PLoS One. 2015 Nov 18;10(11):e0140268. doi: 10.1371/journal.pone.0140268.

eCollection 2015.

Transcriptator: An Automated Computational Pipeline to Annotate Assembled Reads

and Identify Non Coding RNA.

Tripathi KP(1), Evangelista D(1), Zuccaro A(1), Guarracino MR(1).

Author information:

(1)Laboratory for Genomics, Transcriptomics and Proteomics (LAB-GTP), High

Performance Computing and Networking Institute (ICAR), National Research Council

of Italy (CNR), Via Pietro Castellino, 111, Napoli, Italy.

RNA-seq is a new tool to measure RNA transcript counts, using high-throughput

sequencing at an extraordinary accuracy. It provides quantitative means to

explore the transcriptome of an organism of interest. However, interpreting this

extremely large data into biological knowledge is a problem, and

biologist-friendly tools are lacking. In our lab, we developed Transcriptator, a

web application based on a computational Python pipeline with a user-friendly

Java interface. This pipeline uses the web services available for BLAST (Basis

Local Search Alignment Tool), QuickGO and DAVID (Database for Annotation,

Visualization and Integrated Discovery) tools. It offers a report on statistical

analysis of functional and Gene Ontology (GO) annotation's enrichment. It helps

users to identify enriched biological themes, particularly GO terms, pathways,

domains, gene/proteins features and protein-protein interactions related

informations. It clusters the transcripts based on functional annotations and

generates a tabular report for functional and gene ontology annotations for each

submitted transcript to the web server. The implementation of QuickGo

web-services in our pipeline enable the users to carry out GO-Slim analysis,

whereas the integration of PORTRAIT (Prediction of transcriptomic non coding RNA

(ncRNA) by ab initio methods) helps to identify the non coding RNAs and their

regulatory role in transcriptome. In summary, Transcriptator is a useful software

for both NGS and array data. It helps the users to characterize the de-novo

assembled reads, obtained from NGS experiments for non-referenced organisms,

while it also performs the functional enrichment analysis of differentially

expressed transcripts/genes for both RNA-seq and micro-array experiments. It

generates easy to read tables and interactive charts for better understanding of

the data. The pipeline is modular in nature, and provides an opportunity to add

new plugins in the future. Web application is freely available at:

http://www-labgtp.na.icar.cnr.it/Transcriptator.

DOI: 10.1371/journal.pone.0140268

PMCID: PMC4651556

PMID: 26581084 [Indexed for MEDLINE]

402. Database (Oxford). 2015 Nov 13;2015. pii: bav108. doi: 10.1093/database/bav108.

Print 2015.

MAPanalyzer: a novel online tool for analyzing microtubule-associated proteins.

Zhou Y(1), Yang S(1), Mao T(2), Zhang Z(3).

Author information:

(1)State Key Laboratory of Agrobiotechnology and. (2)State Key Laboratory of

Plant Physiology and Biochemistry, College of Biological Sciences, China

Agricultural University, Beijing 100193, China. (3)State Key Laboratory of Plant

Physiology and Biochemistry, College of Biological Sciences, China Agricultural

University, Beijing 100193, China zidingzhang@cau.edu.cn.

The wide functional impacts of microtubules are unleashed and controlled by a

battery of microtubule-associated proteins (MAPs). Specialists in the field

appreciate the diversity of known MAPs and propel the identifications of novel

MAPs. By contrast, there is neither specific database to record known MAPs, nor

MAP predictor that can facilitate the discovery of potential MAPs. We here report

the establishment of a MAP-centered online analysis tool MAPanalyzer, which

consists of a MAP database and a MAP predictor. In the database, a core MAP

dataset, which is fully manually curated from the literature, is further enriched

by MAP information collected via automated pipeline. The core dataset, on the

other hand, enables the building of a novel MAP predictor which combines

specialized machine learning classifiers and the BLAST homology searching tool.

Benchmarks on the curated testing dataset and the Arabidopsis thaliana whole

genome dataset have shown that the proposed predictor outperforms not only its

own components (i.e. the machine learning classifiers and BLAST), but also

another popular homology searching tool, PSI-BLAST. Therefore, MAPanalyzer will

serve as a promising computational resource for the investigations of MAPs.

Database URL: http://systbio.cau.edu.cn/mappred/.

© The Author(s) 2015. Published by Oxford University Press.

DOI: 10.1093/database/bav108

PMCID: PMC4644220

PMID: 26568329 [Indexed for MEDLINE]

403. Source Code Biol Med. 2015 Nov 11;10:11. doi: 10.1186/s13029-015-0043-5.

eCollection 2015.

CyNetworkBMA: a Cytoscape app for inferring gene regulatory networks.

Fronczuk M(1), Raftery AE(2), Yeung KY(1).

Author information:

(1)Institute of Technology, University of Washington, Tacoma, 98402 WA USA.

(2)Department of Statistics, University of Washington, Seattle, 98195 WA USA.

BACKGROUND: Inference of gene networks from expression data is an important

problem in computational biology. Many algorithms have been proposed for solving

the problem efficiently. However, many of the available implementations are

programming libraries that require users to write code, which limits their

accessibility.

RESULTS: We have developed a tool called CyNetworkBMA for inferring gene networks

from expression data that integrates with Cytoscape. Our application offers a

graphical user interface for networkBMA, an efficient implementation of Bayesian

Model Averaging methods for network construction. The client-server architecture

of CyNetworkBMA makes it possible to distribute or centralize computation

depending on user needs.

CONCLUSIONS: CyNetworkBMA is an easy-to-use tool that makes network inference

accessible to non-programmers through seamless integration with Cytoscape.

CyNetworkBMA is available on the Cytoscape App Store at

http://apps.cytoscape.org/apps/cynetworkbma.

DOI: 10.1186/s13029-015-0043-5

PMCID: PMC4642660

PMID: 26566394

404. PLoS One. 2015 Nov 10;10(11):e0137859. doi: 10.1371/journal.pone.0137859.

eCollection 2015.

RNA Thermodynamic Structural Entropy.

Garcia-Martin JA(1), Clote P(1).

Author information:

(1)Department of Biology, Boston College, Chestnut Hill, MA 02467, United States

of America.

Conformational entropy for atomic-level, three dimensional biomolecules is known

experimentally to play an important role in protein-ligand discrimination, yet

reliable computation of entropy remains a difficult problem. Here we describe the

first two accurate and efficient algorithms to compute the conformational entropy

for RNA secondary structures, with respect to the Turner energy model, where free

energy parameters are determined from UV absorption experiments. An algorithm to

compute the derivational entropy for RNA secondary structures had previously been

introduced, using stochastic context free grammars (SCFGs). However, the

numerical value of derivational entropy depends heavily on the chosen context

free grammar and on the training set used to estimate rule probabilities. Using

data from the Rfam database, we determine that both of our thermodynamic methods,

which agree in numerical value, are substantially faster than the SCFG method.

Thermodynamic structural entropy is much smaller than derivational entropy, and

the correlation between length-normalized thermodynamic entropy and derivational

entropy is moderately weak to poor. In applications, we plot the structural

entropy as a function of temperature for known thermoswitches, such as the

repression of heat shock gene expression (ROSE) element, we determine that the

correlation between hammerhead ribozyme cleavage activity and total free energy

is improved by including an additional free energy term arising from

conformational entropy, and we plot the structural entropy of windows of the

HIV-1 genome. Our software RNAentropy can compute structural entropy for any

user-specified temperature, and supports both the Turner'99 and Turner'04 energy

parameters. It follows that RNAentropy is state-of-the-art software to compute

RNA secondary structure conformational entropy. Source code is available at

https://github.com/clotelab/RNAentropy/; a full web server is available at

http://bioinformatics.bc.edu/clotelab/RNAentropy, including source code and

ancillary programs.

DOI: 10.1371/journal.pone.0137859

PMCID: PMC4640541

PMID: 26555444 [Indexed for MEDLINE]

405. Sci Rep. 2015 Nov 6;5:16332. doi: 10.1038/srep16332.

De novo protein conformational sampling using a probabilistic graphical model.

Bhattacharya D(1), Cheng J(1,)(2,)(3).

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, MO 65211,

USA. (2)Informatics Institute, University of Missouri, Columbia, MO 65211, USA.

(3)Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA.

Efficient exploration of protein conformational space remains challenging

especially for large proteins when assembling discretized structural fragments

extracted from a protein structure data database. We propose a fragment-free

probabilistic graphical model, FUSION, for conformational sampling in continuous

space and assess its accuracy using 'blind' protein targets with a length up to

250 residues from the CASP11 structure prediction exercise. The method reduces

sampling bottlenecks, exhibits strong convergence, and demonstrates better

performance than the popular fragment assembly method, ROSETTA, on relatively

larger proteins with a length of more than 150 residues in our benchmark set.

FUSION is freely available through a web server at

http://protein.rnet.missouri.edu/FUSION/.

DOI: 10.1038/srep16332

PMCID: PMC4635387

PMID: 26541939 [Indexed for MEDLINE]

406. Structure. 2015 Nov 3;23(11):2162-70. doi: 10.1016/j.str.2015.08.018. Epub 2015

Oct 9.

ConTemplate Suggests Possible Alternative Conformations for a Query Protein of

Known Structure.

Narunsky A(1), Nepomnyachiy S(2), Ashkenazy H(3), Kolodny R(4), Ben-Tal N(5).

Author information:

(1)Department of Biochemistry and Molecular Biology, George S. Wise Faculty of

Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel. (2)Department of

Computer Science and Engineering, Polytechnic Institute of New York University,

Brooklyn, NY 11201, USA. (3)The Department of Cell Research and Immunology,

George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978,

Israel. (4)Department of Computer Science, University of Haifa, Mount Carmel,

Haifa 31905, Israel. Electronic address: trachel@cs.haifa.ac.il. (5)Department of

Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel

Aviv University, Ramat Aviv 69978, Israel. Electronic address:

bental@ashtoret.tau.ac.il.

Protein function involves conformational changes, but often, for a given protein,

only some of these conformations are known. The missing conformations could be

predicted using the wealth of data in the PDB. Most PDB proteins have multiple

structures, and proteins sharing one similar conformation often share others as

well. The ConTemplate web server (http://bental.tau.ac.il/contemplate) exploits

these observations to suggest conformations for a query protein with at least one

known conformation (or model thereof). We demonstrate ConTemplate on a

ribose-binding protein that undergoes significant conformational changes upon

substrate binding. Querying ConTemplate with the ligand-free (or bound) structure

of the protein produces the ligand-bound (or free) conformation with a

root-mean-square deviation of 1.7 Å (or 2.2 Å); the models are derived from

conformations of other sugar-binding proteins, sharing approximately 30% sequence

identity with the query. The calculation also suggests intermediate conformations

and a pathway between the bound and free conformations.

Copyright © 2015 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.str.2015.08.018

PMID: 26455800 [Indexed for MEDLINE]

407. BMC Res Notes. 2015 Nov 2;8:628. doi: 10.1186/s13104-015-1622-x.

WeBIAS: a web server for publishing bioinformatics applications.

Daniluk P(1,)(2), Wilczyński B(3), Lesyng B(4,)(5).

Author information:

(1)Bioinformatics Laboratory, Mossakowski Medical Research Centre, Pawińskiego 5,

02-106, Warsaw, Poland. pawel@bioexploratorium.pl. (2)Department of Biophysics,

Faculty of Physics, University of Warsaw, Warsaw, Poland.

pawel@bioexploratorium.pl. (3)Institute of Informatics, University of Warsaw,

Warsaw, Poland. bartek@mimuw.edu.pl. (4)Bioinformatics Laboratory, Mossakowski

Medical Research Centre, Pawińskiego 5, 02-106, Warsaw, Poland.

lesyng@icm.edu.pl. (5)Department of Biophysics, Faculty of Physics, University of

Warsaw, Warsaw, Poland. lesyng@icm.edu.pl.

BACKGROUND: One of the requirements for a successful scientific tool is its

availability. Developing a functional web service, however, is usually considered

a mundane and ungratifying task, and quite often neglected. When publishing

bioinformatic applications, such attitude puts additional burden on the reviewers

who have to cope with poorly designed interfaces in order to assess quality of

presented methods, as well as impairs actual usefulness to the scientific

community at large.

RESULTS: In this note we present WeBIAS-a simple, self-contained solution to make

command-line programs accessible through web forms. It comprises a web portal

capable of serving several applications and backend schedulers which carry out

computations. The server handles user registration and authentication, stores

queries and results, and provides a convenient administrator interface. WeBIAS is

implemented in Python and available under GNU Affero General Public License. It

has been developed and tested on GNU/Linux compatible platforms covering a vast

majority of operational WWW servers. Since it is written in pure Python, it

should be easy to deploy also on all other platforms supporting Python (e.g.

Windows, Mac OS X). Documentation and source code, as well as a demonstration

site are available at http://bioinfo.imdik.pan.pl/webias .

CONCLUSIONS: WeBIAS has been designed specifically with ease of installation and

deployment of services in mind. Setting up a simple application requires minimal

effort, yet it is possible to create visually appealing, feature-rich interfaces

for query submission and presentation of results.

DOI: 10.1186/s13104-015-1622-x

PMCID: PMC4629404

PMID: 26526344 [Indexed for MEDLINE]

408. Genome Biol. 2015 Nov 2;16:218. doi: 10.1186/s13059-015-0784-0.

WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system.

Wong N(1,)(2), Liu W(2), Wang X(3,)(4).

Author information:

(1)Department of Biomedical Engineering, Washington University, St Louis, MO,

63130, USA. (2)Department of Radiation Oncology, Washington University School of

Medicine, St Louis, MO, 63108, USA. (3)Department of Biomedical Engineering,

Washington University, St Louis, MO, 63130, USA. xwang@radonc.wustl.edu.

(4)Department of Radiation Oncology, Washington University School of Medicine, St

Louis, MO, 63108, USA. xwang@radonc.wustl.edu.

The CRISPR/Cas9 system has been rapidly adopted for genome editing. However, one

major issue with this system is the lack of robust bioinformatics tools for

design of single guide RNA (sgRNA), which determines the efficacy and specificity

of genome editing. To address this pressing need, we analyze CRISPR RNA-seq data

and identify many novel features that are characteristic of highly potent sgRNAs.

These features are used to develop a bioinformatics tool for genome-wide design

of sgRNAs with improved efficiency. These sgRNAs as well as the design tool are

freely accessible via a web server, WU-CRISPR ( http://crispr.wustl.edu ).

DOI: 10.1186/s13059-015-0784-0

PMCID: PMC4629399

PMID: 26521937 [Indexed for MEDLINE]

409. Bioinformatics. 2015 Nov 1;31(21):3506-13. doi: 10.1093/bioinformatics/btv472.

Epub 2015 Aug 14.

Protein contact prediction by integrating joint evolutionary coupling analysis

and supervised learning.

Ma J(1), Wang S(1), Wang Z(1), Xu J(1).

Author information:

(1)Toyota Technological Institute at Chicago, 6045 S. Kenwood Ave. Chicago,

Illinois 60637 USA.

MOTIVATION: Protein contact prediction is important for protein structure and

functional study. Both evolutionary coupling (EC) analysis and supervised machine

learning methods have been developed, making use of different information

sources. However, contact prediction is still challenging especially for proteins

without a large number of sequence homologs.

RESULTS: This article presents a group graphical lasso (GGL) method for contact

prediction that integrates joint multi-family EC analysis and supervised learning

to improve accuracy on proteins without many sequence homologs. Different from

existing single-family EC analysis that uses residue coevolution information in

only the target protein family, our joint EC analysis uses residue coevolution in

both the target family and its related families, which may have divergent

sequences but similar folds. To implement this, we model a set of related protein

families using Gaussian graphical models and then coestimate their parameters by

maximum-likelihood, subject to the constraint that these parameters shall be

similar to some degree. Our GGL method can also integrate supervised learning

methods to further improve accuracy. Experiments show that our method outperforms

existing methods on proteins without thousands of sequence homologs, and that our

method performs better on both conserved and family-specific contacts.

AVAILABILITY AND IMPLEMENTATION: See http://raptorx.uchicago.edu/ContactMap/ for

a web server implementing the method.

CONTACT: j3xu@ttic.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv472

PMCID: PMC4838177

PMID: 26275894 [Indexed for MEDLINE]

410. Bioinformatics. 2015 Nov 1;31(21):3499-505. doi: 10.1093/bioinformatics/btv390.

Epub 2015 Jun 30.

GDFuzz3D: a method for protein 3D structure reconstruction from contact maps,

based on a non-Euclidean distance function.

Pietal MJ(1), Bujnicki JM(2), Kozlowski LP(3).

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology in Warsaw, Warsaw, Poland, Laboratory of Functional

and Structural Genomics, Centre of New Technologies, University of Warsaw,

Warsaw, Poland and. (2)Laboratory of Bioinformatics and Protein Engineering,

International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland,

Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Adam

Mickiewicz University, Poznan, Poland. (3)Laboratory of Bioinformatics and

Protein Engineering, International Institute of Molecular and Cell Biology in

Warsaw, Warsaw, Poland.

MOTIVATION: To date, only a few distinct successful approaches have been

introduced to reconstruct a protein 3D structure from a map of contacts between

its amino acid residues (a 2D contact map). Current algorithms can infer

structures from information-rich contact maps that contain a limited fraction of

erroneous predictions. However, it is difficult to reconstruct 3D structures from

predicted contact maps that usually contain a high fraction of false contacts.

RESULTS: We describe a new, multi-step protocol that predicts protein 3D

structures from the predicted contact maps. The method is based on a novel

distance function acting on a fuzzy residue proximity graph, which predicts a 2D

distance map from a 2D predicted contact map. The application of a

Multi-Dimensional Scaling algorithm transforms that predicted 2D distance map

into a coarse 3D model, which is further refined by typical modeling programs

into an all-atom representation. We tested our approach on contact maps predicted

de novo by MULTICOM, the top contact map predictor according to CASP10. We show

that our method outperforms FT-COMAR, the state-of-the-art method for 3D

structure reconstruction from 2D maps. For all predicted 2D contact maps of

relatively low sensitivity (60-84%), GDFuzz3D generates more accurate 3D models,

with the average improvement of 4.87 Å in terms of RMSD.

AVAILABILITY AND IMPLEMENTATION: GDFuzz3D server and standalone version are

freely available at http://iimcb.genesilico.pl/gdserver/GDFuzz3D/.

CONTACT: iamb@genesilico.pl

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv390

PMID: 26130575 [Indexed for MEDLINE]

411. Biopolymers. 2015 Nov;104(6):753-63. doi: 10.1002/bip.22703.

AVP-IC50 Pred: Multiple machine learning techniques-based prediction of peptide

antiviral activity in terms of half maximal inhibitory concentration (IC50).

Qureshi A(1), Tandon H(1), Kumar M(1).

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Council of

Scientific and Industrial Research, Sector 39-A, Chandigarh, 160036, India.

Peptide-based antiviral therapeutics has gradually paved their way into

mainstream drug discovery research. Experimental determination of peptides'

antiviral activity as expressed by their IC50 values involves a lot of effort.

Therefore, we have developed "AVP-IC50 Pred," a regression-based algorithm to

predict the antiviral activity in terms of IC50 values (μM). A total of 759

non-redundant peptides from AVPdb and HIPdb were divided into a training/test set

having 683 peptides (T(683)) and a validation set with 76 independent peptides

(V(76)) for evaluation. We utilized important peptide sequence features like

amino-acid compositions, binary profile of N8-C8 residues, physicochemical

properties and their hybrids. Four different machine learning techniques (MLTs)

namely Support vector machine, Random Forest, Instance-based classifier, and

K-Star were employed. During 10-fold cross validation, we achieved maximum

Pearson correlation coefficients (PCCs) of 0.66, 0.64, 0.56, 0.55, respectively,

for the above MLTs using the best combination of feature sets. All the predictive

models also performed well on the independent validation dataset and achieved

maximum PCCs of 0.74, 0.68, 0.59, 0.57, respectively, on the best combination of

feature sets. The AVP-IC50 Pred web server is anticipated to assist the

researchers working on antiviral therapeutics by enabling them to computationally

screen many compounds and focus experimental validation on the most promising set

of peptides, thus reducing cost and time efforts. The server is available at

http://crdd.osdd.net/servers/ic50avp.

© 2015 Wiley Periodicals, Inc.

DOI: 10.1002/bip.22703

PMID: 26213387 [Indexed for MEDLINE]

412. IEEE/ACM Trans Comput Biol Bioinform. 2015 Nov-Dec;12(6):1385-93. doi:

10.1109/TCBB.2015.2418773.

PRBP: Prediction of RNA-Binding Proteins Using a Random Forest Algorithm Combined

with an RNA-Binding Residue Predictor.

Ma X, Guo J, Xiao K, Sun X.

The prediction of RNA-binding proteins is an incredibly challenging problem in

computational biology. Although great progress has been made using various

machine learning approaches with numerous features, the problem is still far from

being solved. In this study, we attempt to predict RNA-binding proteins directly

from amino acid sequences. A novel approach, PRBP predicts RNA-binding proteins

using the information of predicted RNA-binding residues in conjunction with a

random forest based method. For a given protein, we first predict its RNA-binding

residues and then judge whether the protein binds RNA or not based on information

from that prediction. If the protein cannot be identified by the information

associated with its predicted RNA-binding residues, then a novel random forest

predictor is used to determine if the query protein is a RNA-binding protein. We

incorporated features of evolutionary information combined with physicochemical

features (EIPP) and amino acid composition feature to establish the random forest

predictor. Feature analysis showed that EIPP contributed the most to the

prediction of RNA-binding proteins. The results also showed that the information

from the RNA-binding residue prediction improved the overall performance of our

RNA-binding protein prediction. It is anticipated that the PRBP method will

become a useful tool for identifying RNA-binding proteins. A PRBP Web server

implementation is freely available at http://www.cbi.seu.edu.cn/PRBP/.

DOI: 10.1109/TCBB.2015.2418773

PMID: 26671809 [Indexed for MEDLINE]

413. J Mol Graph Model. 2015 Nov;62:253-61. doi: 10.1016/j.jmgm.2015.10.007. Epub 2015

Oct 19.

Vienna Soil-Organic-Matter Modeler--Generating condensed-phase models of humic

substances.

Sündermann A(1), Solc R(2), Tunega D(2), Haberhauer G(2), Gerzabek MH(2),

Oostenbrink C(3).

Author information:

(1)Department of Material Sciences and Process Engineering, Institute of

Molecular Modeling and Simulation, University of Natural Resources and Life

Sciences Vienna, Muthgasse 18, A-1190 Vienna, Austria. (2)Department of Forest-

and Soil Sciences, Institute of Soil Research, University of Natural Resources

and Life Sciences Vienna, Peter-Jordan-Straße 82, A-1190 Vienna, Austria.

(3)Department of Material Sciences and Process Engineering, Institute of

Molecular Modeling and Simulation, University of Natural Resources and Life

Sciences Vienna, Muthgasse 18, A-1190 Vienna, Austria. Electronic address:

chris.oostenbrink@boku.ac.at.

Humic substances are ubiquitous in the environment and have manifold functions.

While their composition is well known, information on the chemical structure and

three-dimensional conformation is scarce. Here we describe the Vienna

Soil-Organic-Matter Modeler, which is an online tool to generate condensed phase

computer models of humic substances (http://somm.boku.ac.at). Many different

models can be created that reflect the diversity in composition and conformations

of the constituting molecules. To exemplify the modeler, 18 different models are

generated based on two experimentally determined compositions, to explicitly

study the effect of varying e.g. the amount of water molecules in the models or

the pH. Molecular dynamics simulations were performed on the models, which were

subsequently analyzed in terms of structure, interactions and dynamics, linking

macroscopic observables to the microscopic composition of the systems. We are

convinced that this new tool opens the way for a wide range of in silico studies

on soil organic matter.

Copyright © 2015 The Authors. Published by Elsevier Inc. All rights reserved.

DOI: 10.1016/j.jmgm.2015.10.007

PMID: 26521208 [Indexed for MEDLINE]

414. Mol Inform. 2015 Nov;34(11-12):761-70. doi: 10.1002/minf.201500031. Epub 2015 Sep

14.

Improving tRNAscan-SE Annotation Results via Ensemble Classifiers.

Zou Q(1,)(2), Guo J(1), Ju Y(1), Wu M(1), Zeng X(1), Hong Z(3).

Author information:

(1)School of Information Science and Technology, Xiamen University, Xiamen

361005, China. (2)School of Computer Science and Technology, Tianjin University,

Tianjin 300072, China. (3)Software School, Xiamen University, Xiamen 361005,

China. hongzl@xmu.edu.cn.

tRNAScan-SE is a tRNA detection program that is widely used for tRNA annotation;

however, the false positive rate of tRNAScan-SE is unacceptable for large

sequences. Here, we used a machine learning method to try to improve the

tRNAScan-SE results. A new predictor, tRNA-Predict, was designed. We obtained

real and pseudo-tRNA sequences as training data sets using tRNAScan-SE and

constructed three different tRNA feature sets. We then set up an ensemble

classifier, LibMutil, to predict tRNAs from the training data. The positive data

set of 623 tRNA sequences was obtained from tRNAdb 2009 and the negative data set

was the false positive tRNAs predicted by tRNAscan-SE. Our in silico experiments

revealed a prediction accuracy rate of 95.1 % for tRNA-Predict using 10-fold

cross-validation. tRNA-Predict was developed to distinguish functional tRNAs from

pseudo-tRNAs rather than to predict tRNAs from a genome-wide scan. However,

tRNA-Predict can work with the output of tRNAscan-SE, which is a genome-wide

scanning method, to improve the tRNAscan-SE annotation results. The tRNA-Predict

web server is accessible at http://datamining.xmu.edu.cn/∼gjs/tRNA-Predict.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

DOI: 10.1002/minf.201500031

PMID: 27491037

415. BMC Bioinformatics. 2015 Oct 23;16:339. doi: 10.1186/s12859-015-0776-9.

PDB-Explorer: a web-based interactive map of the protein data bank in shape

space.

Jin X(1), Awale M(2), Zasso M(3), Kostro D(4), Patiny L(5), Reymond JL(6).

Author information:

(1)Department of Chemistry and Biochemistry, University of Berne, Freiestrasse 3,

3012, Berne, Switzerland. xian.jin@dcb.unibe.ch. (2)Department of Chemistry and

Biochemistry, University of Berne, Freiestrasse 3, 3012, Berne, Switzerland.

mahendra.awale@dcb.unibe.ch. (3)Ecole Polytechnique Fédérale de Lausanne (EPFL),

Institute of Chemical Sciences and Engineering (ISIC), Lausanne, 1015,

Switzerland. mic.besace@gmail.com. (4)Ecole Polytechnique Fédérale de Lausanne

(EPFL), Institute of Chemical Sciences and Engineering (ISIC), Lausanne, 1015,

Switzerland. kostro.d@gmail.com. (5)Ecole Polytechnique Fédérale de Lausanne

(EPFL), Institute of Chemical Sciences and Engineering (ISIC), Lausanne, 1015,

Switzerland. luc.patiny@epfl.ch. (6)Department of Chemistry and Biochemistry,

University of Berne, Freiestrasse 3, 3012, Berne, Switzerland.

jean-louis.reymond@dcb.unibe.ch.

BACKGROUND: The RCSB Protein Data Bank (PDB) provides public access to

experimentally determined 3D-structures of biological macromolecules (proteins,

peptides and nucleic acids). While various tools are available to explore the

PDB, options to access the global structural diversity of the entire PDB and to

perceive relationships between PDB structures remain very limited.

METHODS: A 136-dimensional atom pair 3D-fingerprint for proteins (3DP) counting

categorized atom pairs at increasing through-space distances was designed to

represent the molecular shape of PDB-entries. Nearest neighbor searches examples

were reported exemplifying the ability of 3DP-similarity to identify closely

related biomolecules from small peptides to enzyme and large multiprotein

complexes such as virus particles. The principle component analysis was used to

obtain the visualization of PDB in 3DP-space.

RESULTS: The 3DP property space groups proteins and protein assemblies according

to their 3D-shape similarity, yet shows exquisite ability to distinguish between

closely related structures. An interactive website called PDB-Explorer is

presented featuring a color-coded interactive map of PDB in 3DP-space. Each pixel

of the map contains one or more PDB-entries which are directly visualized as

ribbon diagrams when the pixel is selected. The PDB-Explorer website allows

performing 3DP-nearest neighbor searches of any PDB-entry or of any structure

uploaded as protein-type PDB file. All functionalities on the website are

implemented in JavaScript in a platform-independent manner and draw data from a

server that is updated daily with the latest PDB additions, ensuring complete and

up-to-date coverage. The essentially instantaneous 3DP-similarity search with the

PDB-Explorer provides results comparable to those of much slower 3D-alignment

algorithms, and automatically clusters proteins from the same superfamilies in

tight groups.

CONCLUSION: A chemical space classification of PDB based on molecular shape was

obtained using a new atom-pair 3D-fingerprint for proteins and implemented in a

web-based database exploration tool comprising an interactive color-coded map of

the PDB chemical space and a nearest neighbor search tool. The PDB-Explorer

website is freely available at www.cheminfo.org/pdbexplorer and represents an

unprecedented opportunity to interactively visualize and explore the structural

diversity of the PDB. ᅟ

DOI: 10.1186/s12859-015-0776-9

PMCID: PMC4619230

PMID: 26493835 [Indexed for MEDLINE]

416. BMC Bioinformatics. 2015 Oct 23;16:337. doi: 10.1186/s12859-015-0775-x.

A large-scale conformation sampling and evaluation server for protein tertiary

structure prediction and its assessment in CASP11.

Li J(1), Cao R(2), Cheng J(3,)(4,)(5).

Author information:

(1)Department of Computer Science, University of Missouri-Columbia, Columbia, MO,

65211, USA. jl482@mail.missouri.edu. (2)Department of Computer Science,

University of Missouri-Columbia, Columbia, MO, 65211, USA.

rcrg4@mail.missouri.edu. (3)Department of Computer Science, University of

Missouri-Columbia, Columbia, MO, 65211, USA. chengji@missouri.edu. (4)Informatics

Institute, University of Missouri-Columbia, Columbia, MO, 65211, USA.

chengji@missouri.edu. (5)C. Bond Life Science Center, University of

Missouri-Columbia, Columbia, MO, 65211, USA. chengji@missouri.edu.

BACKGROUND: With more and more protein sequences produced in the genomic era,

predicting protein structures from sequences becomes very important for

elucidating the molecular details and functions of these proteins for biomedical

research. Traditional template-based protein structure prediction methods tend to

focus on identifying the best templates, generating the best alignments, and

applying the best energy function to rank models, which often cannot achieve the

best performance because of the difficulty of obtaining best templates,

alignments, and models.

METHODS: We developed a large-scale conformation sampling and evaluation method

and its servers to improve the reliability and robustness of protein structure

prediction. In the first step, our method used a variety of alignment methods to

sample relevant and complementary templates and to generate alternative and

diverse target-template alignments, used a template and alignment combination

protocol to combine alignments, and used template-based and template-free

modeling methods to generate a pool of conformations for a target protein. In the

second step, it used a large number of protein model quality assessment methods

to evaluate and rank the models in the protein model pool, in conjunction with an

exception handling strategy to deal with any additional failure in model ranking.

RESULTS: The method was implemented as two protein structure prediction servers:

MULTICOM-CONSTRUCT and MULTICOM-CLUSTER that participated in the 11th Critical

Assessment of Techniques for Protein Structure Prediction (CASP11) in 2014. The

two servers were ranked among the best 10 server predictors.

CONCLUSIONS: The good performance of our servers in CASP11 demonstrates the

effectiveness and robustness of the large-scale conformation sampling and

evaluation. The MULTICOM server is available at:

http://sysbio.rnet.missouri.edu/multicom\_cluster/.

DOI: 10.1186/s12859-015-0775-x

PMCID: PMC4619059

PMID: 26493701 [Indexed for MEDLINE]

417. BMC Bioinformatics. 2015 Oct 23;16:335. doi: 10.1186/s12859-015-0771-1.

AlloPred: prediction of allosteric pockets on proteins using normal mode

perturbation analysis.

Greener JG(1), Sternberg MJ(2).

Author information:

(1)Centre for Integrative Systems Biology and Bioinformatics, Department of Life

Sciences, Imperial College London, London, SW7 2AZ, UK.

joe.greener13@imperial.ac.uk. (2)Centre for Integrative Systems Biology and

Bioinformatics, Department of Life Sciences, Imperial College London, London, SW7

2AZ, UK.

BACKGROUND: Despite being hugely important in biological processes, allostery is

poorly understood and no universal mechanism has been discovered. Allosteric

drugs are a largely unexplored prospect with many potential advantages over

orthosteric drugs. Computational methods to predict allosteric sites on proteins

are needed to aid the discovery of allosteric drugs, as well as to advance our

fundamental understanding of allostery.

RESULTS: AlloPred, a novel method to predict allosteric pockets on proteins, was

developed. AlloPred uses perturbation of normal modes alongside pocket

descriptors in a machine learning approach that ranks the pockets on a protein.

AlloPred ranked an allosteric pocket top for 23 out of 40 known allosteric

proteins, showing comparable and complementary performance to two existing

methods. In 28 of 40 cases an allosteric pocket was ranked first or second. The

AlloPred web server, freely available at

http://www.sbg.bio.ic.ac.uk/allopred/home, allows visualisation and analysis of

predictions. The source code and dataset information are also available from this

site.

CONCLUSIONS: Perturbation of normal modes can enhance our ability to predict

allosteric sites on proteins. Computational methods such as AlloPred assist drug

discovery efforts by suggesting sites on proteins for further experimental study.

DOI: 10.1186/s12859-015-0771-1

PMCID: PMC4619270

PMID: 26493317 [Indexed for MEDLINE]

418. PLoS Comput Biol. 2015 Oct 23;11(10):e1004343. doi: 10.1371/journal.pcbi.1004343.

eCollection 2015.

Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template

Homology Modeling.

Meier A(1), Söding J(1).

Author information:

(1)Quantitative and Computational Biology, Max Planck Institute for Biophysical

Chemistry, Göttingen, Germany; Gene Center, Ludwig-Maximilians-Universität

München Munich, Munich, Germany.

Homology modeling predicts the 3D structure of a query protein based on the

sequence alignment with one or more template proteins of known structure. Its

great importance for biological research is owed to its speed, simplicity,

reliability and wide applicability, covering more than half of the residues in

protein sequence space. Although multiple templates have been shown to generally

increase model quality over single templates, the information from multiple

templates has so far been combined using empirically motivated, heuristic

approaches. We present here a rigorous statistical framework for multi-template

homology modeling. First, we find that the query proteins' atomic distance

restraints can be accurately described by two-component Gaussian mixtures. This

insight allowed us to apply the standard laws of probability theory to combine

restraints from multiple templates. Second, we derive theoretically optimal

weights to correct for the redundancy among related templates. Third, a heuristic

template selection strategy is proposed. We improve the average GDT-ha model

quality score by 11% over single template modeling and by 6.5% over a

conventional multi-template approach on a set of 1000 query proteins. Robustness

with respect to wrong constraints is likewise improved. We have integrated our

multi-template modeling approach with the popular MODELLER homology modeling

software in our free HHpred server http://toolkit.tuebingen.mpg.de/hhpred and

also offer open source software for running MODELLER with the new restraints at

https://bitbucket.org/soedinglab/hh-suite.

DOI: 10.1371/journal.pcbi.1004343

PMCID: PMC4619893

PMID: 26496371 [Indexed for MEDLINE]

419. PLoS One. 2015 Oct 23;10(10):e0141105. doi: 10.1371/journal.pone.0141105.

eCollection 2015.

Virtual Pharmacist: A Platform for Pharmacogenomics.

Cheng R(1), Leung RK(2), Chen Y(1), Pan Y(1), Tong Y(1), Li Z(1), Ning L(1), Ling

XB(3), He J(1).

Author information:

(1)Department of Biology, South University of Science and Technology of China,

Shenzhen, China. (2)Division of Genomics and Bioinformatics, The Chinese

University of Hong Kong, Hong Kong, China. (3)Departments of Surgery, Stanford

University, Stanford, California, United States of America.

We present Virtual Pharmacist, a web-based platform that takes common types of

high-throughput data, namely microarray SNP genotyping data, FASTQ and Variant

Call Format (VCF) files as inputs, and reports potential drug responses in terms

of efficacy, dosage and toxicity at one glance. Batch submission facilitates

multivariate analysis or data mining of targeted groups. Individual analysis

consists of a report that is readily comprehensible to patients and practioners

who have basic knowledge in pharmacology, a table that summarizes variants and

potential affected drug response according to the US Food and Drug Administration

pharmacogenomic biomarker labeled drug list and PharmGKB, and visualization of a

gene-drug-target network. Group analysis provides the distribution of the

variants and potential affected drug response of a target group, a sample-gene

variant count table, and a sample-drug count table. Our analysis of genomes from

the 1000 Genome Project underlines the potentially differential drug responses

among different human populations. Even within the same population, the findings

from Watson's genome highlight the importance of personalized medicine. Virtual

Pharmacist can be accessed freely at http://www.sustc-genome.org.cn/vp or

installed as a local web server. The codes and documentation are available at the

GitHub repository (https://github.com/VirtualPharmacist/vp). Administrators can

download the source codes to customize access settings for further development.

DOI: 10.1371/journal.pone.0141105

PMCID: PMC4619711

PMID: 26496198 [Indexed for MEDLINE]

420. Sci Rep. 2015 Oct 20;5:15479. doi: 10.1038/srep15479.

DNA binding protein identification by combining pseudo amino acid composition and

profile-based protein representation.

Liu B(1,)(2), Wang S(1), Wang X(1,)(2).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong, China. (2)Key Laboratory of

Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, Guangdong, China.

DNA-binding proteins play an important role in most cellular processes.

Therefore, it is necessary to develop an efficient predictor for identifying

DNA-binding proteins only based on the sequence information of proteins. The

bottleneck for constructing a useful predictor is to find suitable features

capturing the characteristics of DNA binding proteins. We applied PseAAC to DNA

binding protein identification, and PseAAC was further improved by incorporating

the evolutionary information by using profile-based protein representation.

Finally, Combined with Support Vector Machines (SVMs), a predictor called

iDNAPro-PseAAC was proposed. Experimental results on an updated benchmark dataset

showed that iDNAPro-PseAAC outperformed some state-of-the-art approaches, and it

can achieve stable performance on an independent dataset. By using an ensemble

learning approach to incorporate more negative samples (non-DNA binding proteins)

in the training process, the performance of iDNAPro-PseAAC was further improved.

The web server of iDNAPro-PseAAC is available at

http://bioinformatics.hitsz.edu.cn/iDNAPro-PseAAC/.

DOI: 10.1038/srep15479

PMCID: PMC4611492

PMID: 26482832 [Indexed for MEDLINE]

421. Bioinformatics. 2015 Oct 15;31(20):3368-70. doi: 10.1093/bioinformatics/btv382.

Epub 2015 Jun 22.

MetaQuery: a web server for rapid annotation and quantitative analysis of

specific genes in the human gut microbiome.

Nayfach S(1), Fischbach MA(2), Pollard KS(1).

Author information:

(1)Integrative Program in Quantitative Biology, Gladstone Institutes, and

Division of Biostatistics, University of California San Francisco and.

(2)Department of Bioengineering and Therapeutic Sciences and California Institute

for Quantitative Biosciences, University of California San Francisco, San

Francisco, CA, USA.

Microbiome researchers frequently want to know how abundant a particular

microbial gene or pathway is across different human hosts, including its

association with disease and its co-occurrence with other genes or microbial

taxa. With thousands of publicly available metagenomes, these questions should be

easy to answer. However, computational barriers prevent most researchers from

conducting such analyses. We address this problem with MetaQuery, a web

application for rapid and quantitative analysis of specific genes in the human

gut microbiome. The user inputs one or more query genes, and our software returns

the estimated abundance of these genes across 1267 publicly available fecal

metagenomes from American, European and Chinese individuals. In addition, our

application performs downstream statistical analyses to identify features that

are associated with gene variation, including other query genes (i.e. gene

co-variation), taxa, clinical variables (e.g. inflammatory bowel disease and

diabetes) and average genome size. The speed and accessibility of MetaQuery are a

step toward democratizing metagenomics research, which should allow many

researchers to query the abundance and variation of specific genes in the human

gut microbiome.AVAILABILITY AND IMPLEMENTATION: http://metaquery.docpollard.org.

CONTACT: snayfach@gmail.comS UPPLEMENTARY INFORMATION: Supplementary data are

available at Bioinformatics online.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv382

PMCID: PMC4595903

PMID: 26104745 [Indexed for MEDLINE]

422. Bioinformatics. 2015 Oct 15;31(20):3374-6. doi: 10.1093/bioinformatics/btv369.

Epub 2015 Jun 22.

FoldNucleus: web server for the prediction of RNA and protein folding nuclei from

their 3D structures.

Pereyaslavets LB(1), Sokolovsky IV(1), Galzitskaya OV(1).

Author information:

(1)Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow

Region 142290, Russia.

MOTIVATION: To gain insight into how biopolymers fold as quickly as they do, it

is useful to determine which structural elements limit the rate of RNA/protein

folding.

SUMMARY: We have created a new web server, FoldNucleus. Using this server, it is

possible to calculate the folding nucleus for RNA molecules with known 3D

structures-including pseudoknots, tRNAs, hairpins and ribozymes-and for protein

molecules with known 3D structures, as long as they are smaller than 200 amino

acid residues. Researchers can determine and understand which elements of the

structure limit the folding process for various types of RNAs and protein

molecules. Experimental Ф values for 21 proteins can be found and compared with

those determined by our method:

http://bioinfo.protres.ru/resources/phi\_values.htm.

AVAILABILITY AND IMPLEMENTATION: http://bioinfo.protres.ru/foldnucleus/.

CONTACT: ogalzit@vega.protres.ru.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv369

PMID: 26104744 [Indexed for MEDLINE]

423. Bioinformatics. 2015 Oct 15;31(20):3269-75. doi: 10.1093/bioinformatics/btv367.

Epub 2015 Jun 16.

TPpred3 detects and discriminates mitochondrial and chloroplastic targeting

peptides in eukaryotic proteins.

Savojardo C(1), Martelli PL(1), Fariselli P(2), Casadio R(1).

Author information:

(1)Biocomputing Group, University of Bologna, Department of Biology, 40126

Bologna, Italy and. (2)Biocomputing Group, University of Bologna, Department of

Biology, 40126 Bologna, Italy and Department of Computer Science and Engineering,

University of Bologna, 40127 Bologna, Italy.

MOTIVATION: Molecular recognition of N-terminal targeting peptides is the most

common mechanism controlling the import of nuclear-encoded proteins into

mitochondria and chloroplasts. When experimental information is lacking,

computational methods can annotate targeting peptides, and determine their

cleavage sites for characterizing protein localization, function, and mature

protein sequences. The problem of discriminating mitochondrial from chloroplastic

propeptides is particularly relevant when annotating proteomes of photosynthetic

Eukaryotes, endowed with both types of sequences.

RESULTS: Here, we introduce TPpred3, a computational method that given any

Eukaryotic protein sequence performs three different tasks: (i) the detection of

targeting peptides; (ii) their classification as mitochondrial or chloroplastic

and (iii) the precise localization of the cleavage sites in an organelle-specific

framework. Our implementation is based on our TPpred previously introduced. Here,

we integrate a new N-to-1 Extreme Learning Machine specifically designed for the

classification task (ii). For the last task, we introduce an organelle-specific

Support Vector Machine that exploits sequence motifs retrieved with an extensive

motif-discovery analysis of a large set of mitochondrial and chloroplastic

proteins. We show that TPpred3 outperforms the state-of-the-art methods in all

the three tasks.

AVAILABILITY AND IMPLEMENTATION: The method server and datasets are available at

http://tppred3.biocomp.unibo.it.

CONTACT: gigi@biocomp.unibo.it

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv367

PMID: 26079349 [Indexed for MEDLINE]

424. Bioinformatics. 2015 Oct 15;31(20):3362-4. doi: 10.1093/bioinformatics/btv366.

Epub 2015 Jun 14.

PPUS: a web server to predict PUS-specific pseudouridine sites.

Li YH(1), Zhang G(2), Cui Q(3).

Author information:

(1)Institute of Cardiovascular Sciences, Peking University Health Science Center.

(2)Department of Geriatrics and Gerontology, Beijing Huaxin Hospital, the First

Affiliated Hospital of Tsinghua University and. (3)Department of Biomedical

Informatics, Peking University Health Science Center, Beijing, China.

MOTIVATION: Pseudouridine (Ψ), catalyzed by pseudouridine synthase (PUS), is the

most abundant RNA modification and has important cellular functions. Developing

an algorithm to identify Ψ sites is an important work. And it is better if the

algorithm could assign which PUS modifies the Ψ sites. Here, we developed PPUS

(http://lyh.pkmu.cn/ppus/), the first web server to predict PUS-specific Ψ sites.

PPUS: employed support vector machine as the classifier and used nucleotides

around Ψ sites as the features. Currently, PPUS: could accurately predict new Ψ

sites for PUS1, PUS4 and PUS7 in yeast and PUS4 in human. PPUS: is well designed

and friendly to user.

AVAILABILITY AND IMPLEMENTATION: Our web server is available freely for

non-commercial purposes at: http://lyh.pkmu.cn/ppus/

CONTACT: liyanhui@bjmu.edu.cn or cuiqinghua@hsc.pku.edu.cn.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv366

PMID: 26076723 [Indexed for MEDLINE]

425. Int J Epidemiol. 2015 Oct 8. pii: dyv193. [Epub ahead of print]

ViPAR: a software platform for the Virtual Pooling and Analysis of Research Data.

Carter KW, Francis RW, Carter KW(1), Francis RW(1), Bresnahan M(2), Gissler M(3),

Grønborg TK(4), Gross R(5), Gunnes N(6), Hammond G(1), Hornig M(7), Hultman

CM(8), Huttunen J(9), Langridge A(1), Leonard H(1), Newman S(10), Parner ET(4),

Petersson G(8), Reichenberg A(11), Sandin S(8), Schendel DE(12), Schalkwyk L(10),

Sourander A(13), Steadman C(1), Stoltenberg C(14), Suominen A(15), Surén P(6),

Susser E(2), Sylvester Vethanayagam A(16), Yusof Z(8); International

Collaboration for Autism Registry Epidemiology.

Author information:

(1)Telethon Kids Institute, University of Western Australia, Perth, WA,

Australia. (2)Department of Epidemiology, Mailman School of Public Health,

Columbia University, New York, NY, USA, New York State Psychiatric Institute, New

York, NY, USA. (3)National Institute for Health and Welfare, Helsinki, Finland,

NHV Nordic School of Public Health, Gothenburg, Sweden. (4)Department of Public

Health, University of Aarhus, Aarhus, Denmark. (5)Division of Psychiatry, Sheba

Medical Center, Tel Hashomer, Israel, Department of Epidemiology and Preventive

Medicine, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv, Israel.

(6)Norwegian Institute of Public Health, Oslo, Norway. (7)Department of

Epidemiology, Mailman School of Public Health, Columbia University, New York, NY,

USA, Center for Infection and Immunity, Mailman School of Public Health, Columbia

University, New York, NY, USA. (8)Karolinska Institutet, Stockholm, Sweden.

(9)Turku University, Turku, Finland. (10)Institute of Psychiatry, King's College

London, London, UK. (11)Department of Psychosis Studies, Institute of Psychiatry,

King's College London, London, UK, Departments of Preventative Medicine and

Psychiatry, Ischan School of Medicine at Mount Sinai, New York, NY, USA.

(12)Department of Public Health, Section for Epidemiology, University of Aarhus,

Aarhus, Denmark, Department of Economics and Business, National Centre for

Register-based Research, University of Aarhus, Aarhus, Denmark, Lundbeck

Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Copenhagen,

Denmark. (13)Child Psychiatry Research Center, Department of Child Psychiatry,

Turku University, Turku, Finland, Turku University Hospital, Turku, Finland.

(14)Norwegian Institute of Public Health, Oslo, Norway, Department of Global

Public Health and Primary Care, University of Bergen, Bergen, Norway.

(15)Department of Child Psychiatry, Turku University, Turku, Finland and.

(16)University of Aarhus, Aarhus, Denmark.

BACKGROUND: Research studies exploring the determinants of disease require

sufficient statistical power to detect meaningful effects. Sample size is often

increased through centralized pooling of disparately located datasets, though

ethical, privacy and data ownership issues can often hamper this process. Methods

that facilitate the sharing of research data that are sympathetic with these

issues and which allow flexible and detailed statistical analyses are therefore

in critical need. We have created a software platform for the Virtual Pooling and

Analysis of Research data (ViPAR), which employs free and open source methods to

provide researchers with a web-based platform to analyse datasets housed in

disparate locations.

METHODS: Database federation permits controlled access to remotely located

datasets from a central location. The Secure Shell protocol allows data to be

securely exchanged between devices over an insecure network. ViPAR combines these

free technologies into a solution that facilitates 'virtual pooling' where data

can be temporarily pooled into computer memory and made available for analysis

without the need for permanent central storage.

RESULTS: Within the ViPAR infrastructure, remote sites manage their own

harmonized research dataset in a database hosted at their site, while a central

server hosts the data federation component and a secure analysis portal. When an

analysis is initiated, requested data are retrieved from each remote site and

virtually pooled at the central site. The data are then analysed by statistical

software and, on completion, results of the analysis are returned to the user and

the virtually pooled data are removed from memory.

CONCLUSIONS: ViPAR is a secure, flexible and powerful analysis platform built on

open source technology that is currently in use by large international consortia,

and is made publicly available at

[http://bioinformatics.childhealthresearch.org.au/software/vipar/].

© The Author 2015. Published by Oxford University Press on behalf of the

International Epidemiological Association.

DOI: 10.1093/ije/dyv193

PMCID: PMC4864874

PMID: 26452388

426. J Theor Biol. 2015 Oct 7;382:223-34. doi: 10.1016/j.jtbi.2015.06.042. Epub 2015

Jul 9.

mLASSO-Hum: A LASSO-based interpretable human-protein subcellular localization

predictor.

Wan S(1), Mak MW(2), Kung SY(3).

Author information:

(1)Department of Electronic and Information Engineering, The Hong Kong

Polytechnic University, Hong Kong SAR, China. Electronic address:

shibiao.wan@connect.polyu.hk. (2)Department of Electronic and Information

Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China.

Electronic address: enmwmak@polyu.edu.hk. (3)Department of Electrical

Engineering, Princeton University, New Jersey, USA. Electronic address:

kung@princeton.edu.

Knowing the subcellular compartments of human proteins is essential to shed light

on the mechanisms of a broad range of human diseases. In computational methods

for protein subcellular localization, knowledge-based methods (especially gene

ontology (GO) based methods) are known to perform better than sequence-based

methods. However, existing GO-based predictors often lack interpretability and

suffer from overfitting due to the high dimensionality of feature vectors. To

address these problems, this paper proposes an interpretable multi-label

predictor, namely mLASSO-Hum, which can yield sparse and interpretable solutions

for large-scale prediction of human protein subcellular localization. By using

the one-vs-rest LASSO-based classifiers, 87 out of more than 8000 GO terms are

found to play more significant roles in determining the subcellular localization.

Based on these 87 essential GO terms, we can decide not only where a protein

resides within a cell, but also why it is located there. To further exploit

information from the remaining GO terms, a method based on the GO hierarchical

information derived from the depth distance of GO terms is proposed. Experimental

results show that mLASSO-Hum performs significantly better than state-of-the-art

predictors. We also found that in addition to the GO terms from the cellular

component category, GO terms from the other two categories also play important

roles in the final classification decisions. For readers' convenience, the

mLASSO-Hum server is available online at

http://bioinfo.eie.polyu.edu.hk/mLASSOHumServer/.

Copyright © 2015 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2015.06.042

PMID: 26164062 [Indexed for MEDLINE]

427. PLoS One. 2015 Oct 5;10(10):e0139486. doi: 10.1371/journal.pone.0139486.

eCollection 2015.

Engineering Proteins for Thermostability with iRDP Web Server.

Panigrahi P(1), Sule M(1), Ghanate A(2), Ramasamy S(1), Suresh CG(1).

Author information:

(1)Division of Biochemical Sciences, CSIR-National Chemical Laboratory, Pune,

Maharashtra, 411008, India. (2)Division of Chemical Engineering and Process

Development, CSIR-National Chemical Laboratory, Pune, Maharashtra, 411008, India.

Engineering protein molecules with desired structure and biological functions has

been an elusive goal. Development of industrially viable proteins with improved

properties such as stability, catalytic activity and altered specificity by

modifying the structure of an existing protein has widely been targeted through

rational protein engineering. Although a range of factors contributing to thermal

stability have been identified and widely researched, the in silico

implementation of these as strategies directed towards enhancement of protein

stability has not yet been explored extensively. A wide range of structural

analysis tools is currently available for in silico protein engineering. However

these tools concentrate on only a limited number of factors or individual protein

structures, resulting in cumbersome and time-consuming analysis. The iRDP web

server presented here provides a unified platform comprising of iCAPS, iStability

and iMutants modules. Each module addresses different facets of effective

rational engineering of proteins aiming towards enhanced stability. While iCAPS

aids in selection of target protein based on factors contributing to structural

stability, iStability uniquely offers in silico implementation of known

thermostabilization strategies in proteins for identification and stability

prediction of potential stabilizing mutation sites. iMutants aims to assess

mutants based on changes in local interaction network and degree of residue

conservation at the mutation sites. Each module was validated using an

extensively diverse dataset. The server is freely accessible at

http://irdp.ncl.res.in and has no login requirements.

DOI: 10.1371/journal.pone.0139486

PMCID: PMC4593602

PMID: 26436543 [Indexed for MEDLINE]

428. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):321-31. doi:

10.1016/j.gpb.2015.08.004. Epub 2015 Nov 10.

CVTree3 Web Server for Whole-genome-based and Alignment-free Prokaryotic

Phylogeny and Taxonomy.

Zuo G(1), Hao B(2).

Author information:

(1)T-Life Research Center, Department of Physics, Fudan University, Shanghai

200433, China. (2)T-Life Research Center, Department of Physics, Fudan

University, Shanghai 200433, China. Electronic address: hao@mail.itp.ac.cn.

A faithful phylogeny and an objective taxonomy for prokaryotes should agree with

each other and ultimately follow the genome data. With the number of sequenced

genomes reaching tens of thousands, both tree inference and detailed comparison

with taxonomy are great challenges. We now provide one solution in the latest

Release 3.0 of the alignment-free and whole-genome-based web server CVTree3. The

server resides in a cluster of 64 cores and is equipped with an interactive,

collapsible, and expandable tree display. It is capable of comparing the tree

branching order with prokaryotic classification at all taxonomic ranks from

domains down to species and strains. CVTree3 allows for inquiry by taxon names

and trial on lineage modifications. In addition, it reports a summary of

monophyletic and non-monophyletic taxa at all ranks as well as produces

print-quality subtree figures. After giving an overview of retrospective

verification of the CVTree approach, the power of the new server is described for

the mega-classification of prokaryotes and determination of taxonomic placement

of some newly-sequenced genomes. A few discrepancies between CVTree and 16S rRNA

analyses are also summarized with regard to possible taxonomic revisions. CVTree3

is freely accessible to all users at http://tlife.fudan.edu.cn/cvtree3/ without

login requirements.

Copyright © 2015 The Authors. Production and hosting by Elsevier Ltd.. All rights

reserved.

DOI: 10.1016/j.gpb.2015.08.004

PMCID: PMC4678791

PMID: 26563468 [Indexed for MEDLINE]

429. Mol Biol Evol. 2015 Oct;32(10):2798-800. doi: 10.1093/molbev/msv150. Epub 2015

Jun 30.

FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny

Inference Program.

Lefort V(1), Desper R(1), Gascuel O(2).

Author information:

(1)Institut de Biologie Computationnelle, LIRMM, UMR 5506: CNRS & Université de

Montpellier, France. (2)Institut de Biologie Computationnelle, LIRMM, UMR 5506:

CNRS & Université de Montpellier, France gascuel@lirmm.fr.

FastME provides distance algorithms to infer phylogenies. FastME is based on

balanced minimum evolution, which is the very principle of Neighbor Joining (NJ).

FastME improves over NJ by performing topological moves using fast, sophisticated

algorithms. The first version of FastME only included Nearest Neighbor

Interchange. The new 2.0 version also includes Subtree Pruning and Regrafting,

while remaining as fast as NJ and providing a number of facilities: Distance

estimation for DNA and proteins with various models and options, bootstrapping,

and parallel computations. FastME is available using several interfaces:

Command-line (to be integrated in pipelines), PHYLIP-like, and a Web server

(http://www.atgc-montpellier.fr/fastme/).

© The Author 2015. Published by Oxford University Press on behalf of the Society

for Molecular Biology and Evolution.

DOI: 10.1093/molbev/msv150

PMCID: PMC4576710

PMID: 26130081 [Indexed for MEDLINE]

430. Mol Inform. 2015 Oct;34(10):698-701. doi: 10.1002/minf.201500040. Epub 2015 Jul

20.

Pred-hERG: A Novel web-Accessible Computational Tool for Predicting Cardiac

Toxicity.

Braga RC(1,)(2), Alves VM(1), Silva MF(1), Muratov E(3), Fourches D(4), Lião

LM(2), Tropsha A(3), Andrade CH(5).

Author information:

(1)Labmol - Laboratory for Molecular Modeling and Drug Design, Faculty of

Pharmacy, Federal University of Goias, Goiania, Goiás 74605-170, Brazil phone:

+55 62 3209-6451; fax: +55 62 3209-6037. (2)Chemistry Institute, Federal

University of Goias, P. O. Box 131, Goiania, Goiás 74001-970, Brazil.

(3)Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal

Chemistry, Eshelman School of Pharmacy, University of North Carolina, Chapel

Hill, NC, 27599, USA. (4)Department of Chemistry, Bioinformatics Research Center,

North Carolina State University, Raleigh, NC, 27695, USA. (5)Labmol - Laboratory

for Molecular Modeling and Drug Design, Faculty of Pharmacy, Federal University

of Goias, Goiania, Goiás 74605-170, Brazil phone: +55 62 3209-6451; fax:

+55 62 3209-6037. carolina@ufg.br.

The blockage of the hERG K(+) channels is closely associated with lethal cardiac

arrhythmia. The notorious ligand promiscuity of this channel earmarked hERG as

one of the most important antitargets to be considered in early stages of drug

development process. Herein we report on the development of an innovative and

freely accessible web server for early identification of putative hERG blockers

and non-blockers in chemical libraries. We have collected the largest publicly

available curated hERG dataset of 5,984 compounds. We succeed in developing

robust and externally predictive binary (CCR≈0.8) and multiclass models

(accuracy≈0.7). These models are available as a web-service freely available for

public at http://labmol.farmacia.ufg.br/predherg/. Three following outcomes are

available for the users: prediction by binary model, prediction by multi-class

model, and the probability maps of atomic contribution. The Pred-hERG will be

continuously updated and upgraded as new information became available.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

DOI: 10.1002/minf.201500040

PMID: 27490970

431. Bioinformation. 2015 Sep 30;11(9):422-5. doi: 10.6026/97320630011422. eCollection

2015.

SynRio: R and Shiny based application platform for cyanobacterial genome

analysis.

Lakshmanan K(1), Peter AP(1), Mohandass S(1), Varadharaj S(1), Lakshmanan U(1),

Dharmar P(1).

Author information:

(1)National Facility for Marine Cyanobacteria ,Sub-Distributed Bioinformatics

Center, Department of Marine Biotechnology, School of Marine Sciences,

Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

SynRio is a Shiny and R based web analysis portal for viewing Synechocystis PCC

6803 genome, a cyanobacterial genome with data analysis capabilities. The web

based user interface is created using R programming language powered by Shiny

package. This web interface helps in creating interactive genome visualization

based on user provided data selection along with selective data download

options.AVAILABILITY: SinRio is available to download freely from Github -

https://github.com/NFMC/SynRio or from http://www.nfmc.res.in/synrio/. In

addition an online version of the platform is also hosted at nfmc.res.in/synrio,

using shiny server (open source edition) installation.

DOI: 10.6026/97320630011422

PMCID: PMC4620618

PMID: 26527850

432. Nucleic Acids Res. 2015 Sep 30;43(17):8135-45. doi: 10.1093/nar/gkv813. Epub 2015

Aug 17.

Optimizing RNA structures by sequence extensions using RNAcop.

Hecker N(1), Christensen-Dalsgaard M(2), Seemann SE(1), Havgaard JH(1), Stadler

PF(3), Hofacker IL(4), Nielsen H(2), Gorodkin J(5).

Author information:

(1)Center for non-coding RNA in Technology and Health, University of Copenhagen,

Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark Department of Veterinary Clinical

and Animal Science, University of Copenhagen, Grønnegårdsvej 3, 1870

Frederiksberg C, Denmark. (2)Center for non-coding RNA in Technology and Health,

University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark

Department of Cellular and Molecular Medicine, Panum Institute, University of

Copenhagen, Bledgamsvej 3, 2200 Copenhagen N, Denmark. (3)Center for non-coding

RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, 1870

Frederiksberg C, Denmark Bioinformatics Group, Department of Computer Science &

IZBI-Interdisciplinary Center for Bioinformatics & LIFE-Leipzig Research Center

for Civilization Diseases, University Leipzig, Härtelstraße 16-18, 04107 Leipzig,

Germany Institute for Theoretical Chemistry, University of Vienna,

Währingerstraße 17, 1090 Wien, Austria Max Planck Institute for Mathematics in

the Sciences, Inselstraße 22, 04103 Leipzig, Germany Santa Fe Institute, 1399

Hyde Park Road, Santa Fe, NM 87501, USA. (4)Center for non-coding RNA in

Technology and Health, University of Copenhagen, Grønnegårdsvej 3, 1870

Frederiksberg C, Denmark Institute for Theoretical Chemistry, University of

Vienna, Währingerstraße 17, 1090 Wien, Austria. (5)Center for non-coding RNA in

Technology and Health, University of Copenhagen, Grønnegårdsvej 3, 1870

Frederiksberg C, Denmark Department of Veterinary Clinical and Animal Science,

University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark

gorodkin@rth.dk.

A key aspect of RNA secondary structure prediction is the identification of novel

functional elements. This is a challenging task because these elements typically

are embedded in longer transcripts where the borders between the element and

flanking regions have to be defined. The flanking sequences impact the folding of

the functional elements both at the level of computational analyses and when the

element is extracted as a transcript for experimental analysis. Here, we analyze

how different flanking region lengths impact folding into a constrained structure

by computing probabilities of folding for different sizes of flanking regions.

Our method, RNAcop (RNA context optimization by probability), is tested on known

and de novo predicted structures. In vitro experiments support the computational

analysis and suggest that for a number of structures, choosing proper lengths of

flanking regions is critical. RNAcop is available as web server and stand-alone

software via http://rth.dk/resources/rnacop.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv813

PMCID: PMC4787817

PMID: 26283181 [Indexed for MEDLINE]

433. BMC Med. 2015 Sep 28;13:243. doi: 10.1186/s12916-015-0476-3.

A tool to make reporting checklists work.

Marušić A(1,)(2).

Author information:

(1)Journal of Global Health, Split, Croatia. ana.marusic@mefst.hr. (2)Department

of Research in Biomedicine and Health, University of Split School of Medicine,

Split, Croatia. ana.marusic@mefst.hr.

Comment on

BMC Med. 2015;13:221.

Although the use of reporting guidelines has been demonstrated to increase the

completeness and transparency of health research published in journals, there is

still a long way to translate their use to the authors at the time where they are

needed - during the actual research process and manuscript writing. An online

tool for writing methodology section of a randomized controlled trial has been

successfully tested in an experimental setting and provides a direction for the

development of writing tools for health research. Writing tools should not

replace original thinking and the excitement of communicating original

discoveries, but make sure that all relevant data are in the manuscript so that

research results can be understood, critically evaluated and used in practice.

Please see related article: http://www.biomedcentral.com/1741-7015/13/221.

DOI: 10.1186/s12916-015-0476-3

PMCID: PMC4585993

PMID: 26412344 [Indexed for MEDLINE]

434. Bioinformatics. 2015 Sep 15;31(18):3060-2. doi: 10.1093/bioinformatics/btv297.

Epub 2015 May 13.

GEO2Enrichr: browser extension and server app to extract gene sets from GEO and

analyze them for biological functions.

Gundersen GW(1), Jones MR(1), Rouillard AD(1), Kou Y(1), Monteiro CD(2), Feldmann

AS(1), Hu KS(1), Ma'ayan A(1).

Author information:

(1)Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine

at Mount Sinai, New York, NY 10029, USA, BD2K-LINCS Data Coordination and

Integration Center, Mount Sinai Knowledge Management Center for Illuminating the

Druggable Genome and 367 Airport Sector, Goiania, Goias 74075, Brazil Department

of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount

Sinai, New York, NY 10029, USA, BD2K-LINCS Data Coordination and Integration

Center, Mount Sinai Knowledge Management Center for Illuminating the Druggable

Genome and 367 Airport Sector, Goiania, Goias 74075, Brazil Department of

Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai,

New York, NY 10029, USA, BD2K-LINCS Data Coordination and Integration Center,

Mount Sinai Knowledge Management Center for Illuminating the Druggable Genome and

367 Airport Sector, Goiania, Goias 74075, Brazil. (2)Department of Pharmacology

and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY

10029, USA, BD2K-LINCS Data Coordination and Integration Center, Mount Sinai

Knowledge Management Center for Illuminating the Druggable Genome and 367 Airport

Sector, Goiania, Goias 74075, Brazil.

MOTIVATION: Identification of differentially expressed genes is an important step

in extracting knowledge from gene expression profiling studies. The raw

expression data from microarray and other high-throughput technologies is

deposited into the Gene Expression Omnibus (GEO) and served as Simple Omnibus

Format in Text (SOFT) files. However, to extract and analyze differentially

expressed genes from GEO requires significant computational skills.

RESULTS: Here we introduce GEO2Enrichr, a browser extension for extracting

differentially expressed gene sets from GEO and analyzing those sets with

Enrichr, an independent gene set enrichment analysis tool containing over 70 000

annotated gene sets organized into 75 gene-set libraries. GEO2Enrichr adds

JavaScript code to GEO web-pages; this code scrapes user selected accession

numbers and metadata, and then, with one click, users can submit this information

to a web-server application that downloads the SOFT files, parses, cleans and

normalizes the data, identifies the differentially expressed genes, and then

pipes the resulting gene lists to Enrichr for downstream functional analysis.

GEO2Enrichr opens a new avenue for adding functionality to major bioinformatics

resources such GEO by integrating tools and resources without the need for a

plug-in architecture. Importantly, GEO2Enrichr helps researchers to quickly

explore hypotheses with little technical overhead, lowering the barrier of entry

for biologists by automating data processing steps needed for knowledge

extraction from the major repository GEO.

AVAILABILITY AND IMPLEMENTATION: GEO2Enrichr is an open source tool, freely

available for installation as browser extensions at the Chrome Web Store and

FireFox Add-ons. Documentation and a browser independent web application can be

found at http://amp.pharm.mssm.edu/g2e/.

CONTACT: avi.maayan@mssm.edu.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv297

PMCID: PMC4607730

PMID: 25971742 [Indexed for MEDLINE]

435. Sci Rep. 2015 Sep 7;5:13859. doi: 10.1038/srep13859.

Identification and analysis of the N(6)-methyladenosine in the Saccharomyces

cerevisiae transcriptome.

Chen W(1,)(2), Tran H(2), Liang Z(3), Lin H(3), Zhang L(2).

Author information:

(1)Department of Physics, School of Sciences, and Center for Genomics and

Computational Biology, Hebei United University, Tangshan 063000, China.

(2)Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, United

States of America. (3)Key Laboratory for Neuro-Information of Ministry of

Education, Center of Bioinformatics, School of Life Science and Technology,

University of Electronic Science and Technology of China, Chengdu 610054, China.

Knowledge of the distribution of N(6)-methyladenosine (m(6)A) is invaluable for

understanding RNA biological functions. However, limitation in experimental

methods impedes the progress towards the identification of m(6)A site. As a

complement of experimental methods, a support vector machine based-method is

proposed to identify m(6)A sites in Saccharomyces cerevisiae genome. In this

model, RNA sequences are encoded by their nucleotide chemical property and

accumulated nucleotide frequency information. It is observed in the jackknife

test that the accuracy achieved by the proposed model in identifying the m(6)A

site was 78.15%. For the convenience of experimental scientists, a web-server for

the proposed model is provided at http://lin.uestc.edu.cn/server/m6Apred.php.

DOI: 10.1038/srep13859

PMCID: PMC4561376

PMID: 26343792 [Indexed for MEDLINE]

436. J Proteome Res. 2015 Sep 4;14(9):3484-91. doi: 10.1021/acs.jproteome.5b00494.

Epub 2015 Aug 11.

Functional Networks of Highest-Connected Splice Isoforms: From The Chromosome 17

Human Proteome Project.

Li HD(1), Menon R(1), Govindarajoo B(1), Panwar B(1), Zhang Y(1), Omenn GS(1),

Guan Y(1).

Author information:

(1)Department of Computational Medicine and Bioinformatics, ‡Department of

Internal Medicine, §Department of Human Genetics and School of Public Health,

∥Department of Electrical Engineering and Computer Science University of Michigan

, Ann Arbor, Michigan 48109, United States.

Alternative splicing allows a single gene to produce multiple transcript-level

splice isoforms from which the translated proteins may show differences in their

expression and function. Identifying the major functional or canonical isoform is

important for understanding gene and protein functions. Identification and

characterization of splice isoforms is a stated goal of the HUPO Human Proteome

Project and of neXtProt. Multiple efforts have catalogued splice isoforms as

"dominant", "principal", or "major" isoforms based on expression or evolutionary

traits. In contrast, we recently proposed highest connected isoforms (HCIs) as a

new class of canonical isoforms that have the strongest interactions in a

functional network and revealed their significantly higher (differential)

transcript-level expression compared to nonhighest connected isoforms (NCIs)

regardless of tissues/cell lines in the mouse. HCIs and their expression behavior

in the human remain unexplored. Here we identified HCIs for 6157 multi-isoform

genes using a human isoform network that we constructed by integrating a large

compendium of heterogeneous genomic data. We present examples for pairs of

transcript isoforms of ABCC3, RBM34, ERBB2, and ANXA7. We found that functional

networks of isoforms of the same gene can show large differences. Interestingly,

differential expression between HCIs and NCIs was also observed in the human on

an independent set of 940 RNA-seq samples across multiple tissues, including

heart, kidney, and liver. Using proteomic data from normal human retina and

placenta, we showed that HCIs are a promising indicator of expressed protein

isoforms exemplified by NUDFB6 and M6PR. Furthermore, we found that a significant

percentage (20%, p = 0.0003) of human and mouse HCIs are homologues, suggesting

their conservation between species. Our identified HCIs expand the repertoire of

canonical isoforms and are expected to facilitate studying main protein products,

understanding gene regulation, and possibly evolution. The network is available

through our web server as a rich resource for investigating isoform functional

relationships (http://guanlab.ccmb.med.umich.edu/hisonet). All MS/MS data were

available at ProteomeXchange Web site (http://www.proteomexchange.org) through

their identifiers (retina: PXD001242, placenta: PXD000754).

DOI: 10.1021/acs.jproteome.5b00494

PMCID: PMC4993635

PMID: 26216192 [Indexed for MEDLINE]

437. BMC Bioinformatics. 2015 Sep 3;16:282. doi: 10.1186/s12859-015-0711-0.

SFESA: a web server for pairwise alignment refinement by secondary structure

shifts.

Tong J(1), Pei J(2), Grishin NV(3,)(4).

Author information:

(1)Department of Biophysics and Department of Biochemistry, University of Texas

Southwestern Medical Center at Dallas, 6001 Forest Park Road, Dallas, TX,

75390-9050, USA. jing.tong@utsouthwestern.edu. (2)Howard Hughes Medical

Institute, University of Texas Southwestern Medical Center at Dallas, 6001 Forest

Park Road, Dallas, TX, 75390-9050, USA. jpei@chop.swmed.edu. (3)Department of

Biophysics and Department of Biochemistry, University of Texas Southwestern

Medical Center at Dallas, 6001 Forest Park Road, Dallas, TX, 75390-9050, USA.

grishin@chop.swmed.edu. (4)Howard Hughes Medical Institute, University of Texas

Southwestern Medical Center at Dallas, 6001 Forest Park Road, Dallas, TX,

75390-9050, USA. grishin@chop.swmed.edu.

BACKGROUND: Protein sequence alignment is essential for a variety of tasks such

as homology modeling and active site prediction. Alignment errors remain the main

cause of low-quality structure models. A bioinformatics tool to refine alignments

is needed to make protein alignments more accurate.

RESULTS: We developed the SFESA web server to refine pairwise protein sequence

alignments. Compared to the previous version of SFESA, which required a set of 3D

coordinates for a protein, the new server will search a sequence database for the

closest homolog with an available 3D structure to be used as a template. For each

alignment block defined by secondary structure elements in the template, SFESA

evaluates alignment variants generated by local shifts and selects the

best-scoring alignment variant. A scoring function that combines the sequence

score of profile-profile comparison and the structure score of template-derived

contact energy is used for evaluation of alignments. PROMALS pairwise alignments

refined by SFESA are more accurate than those produced by current advanced

alignment methods such as HHpred and CNFpred. In addition, SFESA also improves

alignments generated by other software.

CONCLUSIONS: SFESA is a web-based tool for alignment refinement, designed for

researchers to compute, refine, and evaluate pairwise alignments with a combined

sequence and structure scoring of alignment blocks. To our knowledge, the SFESA

web server is the only tool that refines alignments by evaluating local shifts of

secondary structure elements. The SFESA web server is available at

http://prodata.swmed.edu/sfesa.

DOI: 10.1186/s12859-015-0711-0

PMCID: PMC4558796

PMID: 26335387 [Indexed for MEDLINE]

438. J Integr Bioinform. 2015 Sep 3;12(1):257. doi: 10.2390/biecoll-jib-2015-257.

Shared bioinformatics databases within the Unipro UGENE platform.

Protsyuk IV, Grekhov GA, Tiunov AV, Fursov MY.

Unipro UGENE is an open-source bioinformatics toolkit that integrates popular

tools along with original instruments for molecular biologists within a unified

user interface. Nowadays, most bioinformatics desktop applications, including

UGENE, make use of a local data model while processing different types of data.

Such an approach causes an inconvenience for scientists working cooperatively and

relying on the same data. This refers to the need of making multiple copies of

certain files for every workplace and maintaining synchronization between them in

case of modifications. Therefore, we focused on delivering a collaborative work

into the UGENE user experience. Currently, several UGENE installations can be

connected to a designated shared database and users can interact with it

simultaneously. Such databases can be created by UGENE users and be used at their

discretion. Objects of each data type, supported by UGENE such as sequences,

annotations, multiple alignments, etc., can now be easily imported from or

exported to a remote storage. One of the main advantages of this system, compared

to existing ones, is the almost simultaneous access of client applications to

shared data regardless of their volume. Moreover, the system is capable of

storing millions of objects. The storage itself is a regular database server so

even an inexpert user is able to deploy it. Thus, UGENE may provide access to

shared data for users located, for example, in the same laboratory or

institution. UGENE is available at: http://ugene.net/download.html.

DOI: 10.2390/biecoll-jib-2015-257

PMID: 26527191 [Indexed for MEDLINE]

439. PLoS One. 2015 Sep 3;10(9):e0136990. doi: 10.1371/journal.pone.0136990.

eCollection 2015.

AntiAngioPred: A Server for Prediction of Anti-Angiogenic Peptides.

Ettayapuram Ramaprasad AS(1), Singh S(2), Gajendra P S R(2), Venkatesan S(1).

Author information:

(1)Chemical Laboratory, Central Leather Research Institute, Council of Scientific

and Industrial Research, Adyar, Chennai, India. (2)Bioinformatics Centre,

CSIR-Institute of Microbial Technology, Chandigarh, India.

The process of angiogenesis is a vital step towards the formation of malignant

tumors. Anti-angiogenic peptides are therefore promising candidates in the

treatment of cancer. In this study, we have collected anti-angiogenic peptides

from the literature and analyzed the residue preference in these peptides.

Residues like Cys, Pro, Ser, Arg, Trp, Thr and Gly are preferred while Ala, Asp,

Ile, Leu, Val and Phe are not preferred in these peptides. There is a positional

preference of Ser, Pro, Trp and Cys in the N terminal region and Cys, Gly and Arg

in the C terminal region of anti-angiogenic peptides. Motif analysis suggests the

motifs "CG-G", "TC", "SC", "SP-S", etc., which are highly prominent in

anti-angiogenic peptides. Based on the primary analysis, we developed prediction

models using different machine learning based methods. The maximum accuracy and

MCC for amino acid composition based model is 80.9% and 0.62 respectively. The

performance of the models on independent dataset is also reasonable. Based on the

above study, we have developed a user-friendly web server named "AntiAngioPred"

for the prediction of anti-angiogenic peptides. AntiAngioPred web server is

freely accessible at

http://clri.res.in/subramanian/tools/antiangiopred/index.html (mirror site:

http://crdd.osdd.net/raghava/antiangiopred/).

DOI: 10.1371/journal.pone.0136990

PMCID: PMC4559406

PMID: 26335203 [Indexed for MEDLINE]

440. Bioinformatics. 2015 Sep 1;31(17):2897-9. doi: 10.1093/bioinformatics/btv292.

Epub 2015 May 7.

MemGen: a general web server for the setup of lipid membrane simulation systems.

Knight CJ(1), Hub JS(1).

Author information:

(1)Georg-August-University Göttingen, Institute for Microbiology and Genetics,

37077 Göttingen, Germany.

MOTIVATION: Molecular dynamics simulations provide atomic insight into the

physicochemical characteristics of lipid membranes and hence, a wide range of

force field families capable of modelling various lipid types have been developed

in recent years. To model membranes in a biologically realistic lipid

composition, simulation systems containing multiple different lipids must be

assembled.

RESULTS: We present a new web service called MemGen that is capable of setting up

simulation systems of heterogenous lipid membranes. MemGen is not restricted to

certain lipid force fields or lipid types, but instead builds membranes from

uploaded structure files which may contain any kind of amphiphilic molecule.

MemGen works with any all-atom or united-atom lipid representation.

AVAILABILITY AND IMPLEMENTATION: MemGen is freely available without registration

at http://memgen.uni-goettingen.de.

CONTACT: jhub@gwdg.de

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv292

PMID: 25957354 [Indexed for MEDLINE]

441. Bioinformatics. 2015 Sep 1;31(17):2882-4. doi: 10.1093/bioinformatics/btv287.

Epub 2015 May 7.

Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data.

Aßhauer KP(1), Wemheuer B(2), Daniel R(2), Meinicke P(1).

Author information:

(1)Department of Bioinformatics and. (2)Department of Genomic and Applied

Microbiology and Göttingen Genomics Laboratory, Institute of Microbiology and

Genetics, Georg-August-University Göttingen, 37077 Göttingen, Germany.

MOTIVATION: The characterization of phylogenetic and functional diversity is a

key element in the analysis of microbial communities. Amplicon-based sequencing

of marker genes, such as 16S rRNA, is a powerful tool for assessing and comparing

the structure of microbial communities at a high phylogenetic resolution. Because

16S rRNA sequencing is more cost-effective than whole metagenome shotgun

sequencing, marker gene analysis is frequently used for broad studies that

involve a large number of different samples. However, in comparison to shotgun

sequencing approaches, insights into the functional capabilities of the community

get lost when restricting the analysis to taxonomic assignment of 16S rRNA data.

RESULTS: Tax4Fun is a software package that predicts the functional capabilities

of microbial communities based on 16S rRNA datasets. We evaluated Tax4Fun on a

range of paired metagenome/16S rRNA datasets to assess its performance. Our

results indicate that Tax4Fun provides a good approximation to functional

profiles obtained from metagenomic shotgun sequencing approaches.

AVAILABILITY AND IMPLEMENTATION: Tax4Fun is an open-source R package and

applicable to output as obtained from the SILVAngs web server or the application

of QIIME with a SILVA database extension. Tax4Fun is freely available for

download at http://tax4fun.gobics.de/.

CONTACT: kasshau@gwdg.de

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv287

PMCID: PMC4547618

PMID: 25957349 [Indexed for MEDLINE]

442. Bioinformatics. 2015 Sep 1;31(17):2816-21. doi: 10.1093/bioinformatics/btv291.

Epub 2015 May 7.

INPS: predicting the impact of non-synonymous variations on protein stability

from sequence.

Fariselli P(1), Martelli PL(2), Savojardo C(2), Casadio R(2).

Author information:

(1)Biocomputing Group, Department of Biology, University of Bologna, 40126

Bologna and Department of Computer Science and Engineering, University of

Bologna, 40127 Bologna, Italy. (2)Biocomputing Group, Department of Biology,

University of Bologna, 40126 Bologna and.

MOTIVATION: A tool for reliably predicting the impact of variations on protein

stability is extremely important for both protein engineering and for

understanding the effects of Mendelian and somatic mutations in the genome. Next

Generation Sequencing studies are constantly increasing the number of protein

sequences. Given the huge disproportion between protein sequences and structures,

there is a need for tools suited to annotate the effect of mutations starting

from protein sequence without relying on the structure. Here, we describe INPS, a

novel approach for annotating the effect of non-synonymous mutations on the

protein stability from its sequence. INPS is based on SVM regression and it is

trained to predict the thermodynamic free energy change upon single-point

variations in protein sequences.

RESULTS: We show that INPS performs similarly to the state-of-the-art methods

based on protein structure when tested in cross-validation on a non-redundant

dataset. INPS performs very well also on a newly generated dataset consisting of

a number of variations occurring in the tumor suppressor protein p53. Our results

suggest that INPS is a tool suited for computing the effect of non-synonymous

polymorphisms on protein stability when the protein structure is not available.

We also show that INPS predictions are complementary to those of the

state-of-the-art, structure-based method mCSM. When the two methods are combined,

the overall prediction on the p53 set scores significantly higher than those of

the single methods.

AVAILABILITY AND IMPLEMENTATION: The presented method is available as web server

at http://inps.biocomp.unibo.it.

CONTACT: piero.fariselli@unibo.it

SUPPLEMENTARY INFORMATION: Supplementary Materials are available at

Bioinformatics online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv291

PMID: 25957347 [Indexed for MEDLINE]

443. Bioinformatics. 2015 Sep 1;31(17):2808-15. doi: 10.1093/bioinformatics/btv282.

Epub 2015 May 5.

Automated band annotation for RNA structure probing experiments with numerous

capillary electrophoresis profiles.

Lee S(1), Kim H(1), Tian S(2), Lee T(1), Yoon S(3), Das R(4).

Author information:

(1)Department of ECE, Seoul National University, Seoul 151-744, Korea.

(2)Department of Biochemistry, Stanford University School of Medicine, Stanford,

CA 94305, USA. (3)Department of ECE, Seoul National University, Seoul 151-744,

Korea, Interdisciplinary Program in Bionformatics, Seoul National University,

Seoul 151-744, Korea and. (4)Department of Biochemistry, Stanford University

School of Medicine, Stanford, CA 94305, USA, Department of Physics, Stanford

University, Stanford, CA 94305, USA.

MOTIVATION: Capillary electrophoresis (CE) is a powerful approach for structural

analysis of nucleic acids, with recent high-throughput variants enabling

three-dimensional RNA modeling and the discovery of new rules for RNA structure

design. Among the steps composing CE analysis, the process of finding each band

in an electrophoretic trace and mapping it to a position in the nucleic acid

sequence has required significant manual inspection and remains the most

time-consuming and error-prone step. The few available tools seeking to automate

this band annotation have achieved limited accuracy and have not taken advantage

of information across dozens of profiles routinely acquired in high-throughput

measurements.

RESULTS: We present a dynamic-programming-based approach to automate band

annotation for high-throughput capillary electrophoresis. The approach is

uniquely able to define and optimize a robust target function that takes into

account multiple CE profiles (sequencing ladders, different chemical probes,

different mutants) collected for the RNA. Over a large benchmark of multi-profile

datasets for biological RNAs and designed RNAs from the EteRNA project, the

method outperforms prior tools (QuSHAPE and FAST) significantly in terms of

accuracy compared with gold-standard manual annotations. The amount of

computation required is reasonable at a few seconds per dataset. We also

introduce an 'E-score' metric to automatically assess the reliability of the band

annotation and show it to be practically useful in flagging uncertainties in band

annotation for further inspection.

AVAILABILITY AND IMPLEMENTATION: The implementation of the proposed algorithm is

included in the HiTRACE software, freely available as an online server and for

download at http://hitrace.stanford.edu.

CONTACT: sryoon@snu.ac.kr or rhiju@stanford.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv282

PMCID: PMC4560050

PMID: 25943472 [Indexed for MEDLINE]

444. Bioinformatics. 2015 Sep 1;31(17):2794-800. doi: 10.1093/bioinformatics/btv276.

Epub 2015 May 4.

Phylesystem: a git-based data store for community-curated phylogenetic estimates.

McTavish EJ(1), Hinchliff CE(2), Allman JF(3), Brown JW(2), Cranston KA(4),

Holder MT(1), Rees JA(4), Smith SA(2).

Author information:

(1)Department of Ecology and Evolutionary Biology, University of Kansas,

Lawrence, KS, USA, Heidelberg Institute for Theoretical Studies, Heidelberg

69118, Germany. (2)Department of Ecology and Evolutionary Biology, University of

Michigan, Ann Arbor, MI, USA. (3)Interrobang Corporation, Wake Forest, NC 27587,

USA and. (4)National Evolutionary Synthesis Center, Duke University, Durham, NC,

USA.

MOTIVATION: Phylogenetic estimates from published studies can be archived using

general platforms like Dryad (Vision, 2010) or TreeBASE (Sanderson et al., 1994).

Such services fulfill a crucial role in ensuring transparency and reproducibility

in phylogenetic research. However, digital tree data files often require some

editing (e.g. rerooting) to improve the accuracy and reusability of the

phylogenetic statements. Furthermore, establishing the mapping between tip labels

used in a tree and taxa in a single common taxonomy dramatically improves the

ability of other researchers to reuse phylogenetic estimates. As the process of

curating a published phylogenetic estimate is not error-free, retaining a full

record of the provenance of edits to a tree is crucial for openness, allowing

editors to receive credit for their work and making errors introduced during

curation easier to correct.

RESULTS: Here, we report the development of software infrastructure to support

the open curation of phylogenetic data by the community of biologists. The

backend of the system provides an interface for the standard database operations

of creating, reading, updating and deleting records by making commits to a git

repository. The record of the history of edits to a tree is preserved by git's

version control features. Hosting this data store on GitHub (http://github.com/)

provides open access to the data store using tools familiar to many developers.

We have deployed a server running the 'phylesystem-api', which wraps the

interactions with git and GitHub. The Open Tree of Life project has also

developed and deployed a JavaScript application that uses the phylesystem-api and

other web services to enable input and curation of published phylogenetic

statements.

AVAILABILITY AND IMPLEMENTATION: Source code for the web service layer is

available at https://github.com/OpenTreeOfLife/phylesystem-api. The data store

can be cloned from: https://github.com/OpenTreeOfLife/phylesystem. A web

application that uses the phylesystem web services is deployed at

http://tree.opentreeoflife.org/curator. Code for that tool is available from

https://github.com/OpenTreeOfLife/opentree.

CONTACT: mtholder@gmail.com.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv276

PMCID: PMC4547614

PMID: 25940563 [Indexed for MEDLINE]

445. Bioinformatics. 2015 Sep 1;31(17):2891-3. doi: 10.1093/bioinformatics/btv221.

Epub 2015 Apr 24.

iFoldRNA v2: folding RNA with constraints.

Krokhotin A(1), Houlihan K(1), Dokholyan NV(1).

Author information:

(1)Department of Biochemistry & Biophysics, University of North Carolina, Chapel

Hill, NC, 27599, USA.

A key to understanding RNA function is to uncover its complex 3D structure.

Experimental methods used for determining RNA 3D structures are technologically

challenging and laborious, which makes the development of computational

prediction methods of substantial interest. Previously, we developed the iFoldRNA

server that allows accurate prediction of short (<50 nt) tertiary RNA structures

starting from primary sequences. Here, we present a new version of the iFoldRNA

server that permits the prediction of tertiary structure of RNAs as long as a few

hundred nucleotides. This substantial increase in the server capacity is achieved

by utilization of experimental information such as base-pairing and

hydroxyl-radical probing. We demonstrate a significant benefit provided by

integration of experimental data and computational methods.AVAILABILITY AND

IMPLEMENTATION: http://ifoldrna.dokhlab.org

CONTACT: dokh@unc.eu.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv221

PMCID: PMC4547609

PMID: 25910700 [Indexed for MEDLINE]

446. Curr Plant Biol. 2015 Sep-Dec;3-4:48-55.

CressInt: a user-friendly web resource for genome-scale exploration of gene

regulation in Arabidopsis thaliana.

Chen X(1), Ernst K(1), Soman F(2), Borowczak M(1), Weirauch MT(3).

Author information:

(1)Department of Electrical Engineering and Computing Systems, College of

Engineering and Applied Sciences, University of Cincinnati, Cincinnati, OH 45221;

Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital

Medical Center, Department of Pediatrics, College of Medicine, University of

Cincinnati, Cincinnati, OH 45229. (2)Department of Electrical Engineering and

Computing Systems, College of Engineering and Applied Sciences, University of

Cincinnati, Cincinnati, OH 45221. (3)Center for Autoimmune Genomics and Etiology,

Cincinnati Children's Hospital Medical Center, Department of Pediatrics, College

of Medicine, University of Cincinnati, Cincinnati, OH 45229; Division of

Biomedical Informatics and Division of Developmental Biology, Cincinnati

Children's Hospital Medical Center, Department of Pediatrics, College of

Medicine, University of Cincinnati, Cincinnati, OH 45229.

The thale cress Arabidopsis thaliana is a powerful model organism for studying a

wide variety of biological processes. Recent advances in sequencing technology

have resulted in a wealth of information describing numerous aspects of A.

thaliana genome function. However, there is a relative paucity of computational

systems for efficiently and effectively using these data to create testable

hypotheses. We present CressInt, a user-friendly web resource for exploring gene

regulatory mechanisms in A. thaliana on a genomic scale. The CressInt system

incorporates a variety of genome-wide data types relevant to gene regulation,

including transcription factor (TF) binding site models, ChIP-seq, DNase-seq,

eQTLs, and GWAS. We demonstrate the utility of CressInt by showing how the system

can be used to (1) Identify TFs binding to the promoter of a gene of interest;

(2) identify genetic variants that are likely to impact TF binding based on a

ChIP-seq dataset; and (3) identify specific TFs whose binding might be impacted

by phenotype-associated variants. CressInt is freely available at

http://cressint.cchmc.org.

DOI: 10.1016/j.cpb.2015.09.001

PMCID: PMC4740912

PMID: 26855883

447. Genet Med. 2015 Sep;17(9):714-8. doi: 10.1038/gim.2014.180. Epub 2014 Dec 18.

SG-ADVISER CNV: copy-number variant annotation and interpretation.

Erikson GA(1,)(2), Deshpande N(1,)(2), Kesavan BG(1,)(2), Torkamani

A(1,)(2,)(3,)(4).

Author information:

(1)Scripps Health, La Jolla, California, USA. (2)Scripps Translational Science

Institute, La Jolla, California, USA. (3)Department of Integrative Structural and

Computational Biology, Scripps Research Institute, La Jolla, California, USA.

(4)Cypher Genomics, La Jolla, California, USA.

PURPOSE: Copy-number variants have been associated with a variety of diseases,

especially cancer, autism, schizophrenia, and developmental delay. The majority

of clinically relevant events occur de novo, necessitating the interpretation of

novel events. In this light, we present the Scripps Genome ADVISER CNV annotation

pipeline and Web server, which aims to fill the gap between copy number variant

detection and interpretation by performing in-depth annotations and functional

predictions for copy number variants.

METHODS: The Scripps Genome ADVISER CNV suite includes a Web server interface to

a high-performance computing environment for calculations of annotations and a

table-based user interface that allows for the execution of numerous

annotation-based variant filtration strategies and statistics.

RESULTS: The annotation results include details regarding location, impact on the

coding portion of genes, allele frequency information (including allele

frequencies from the Scripps Wellderly cohort), and overlap information with

other reference data sets (including ClinVar, DGV, DECIPHER). A summary variant

classification is produced (ADVISER score) based on the American College of

Medical Genetics and Genomics scoring guidelines. We demonstrate >90%

sensitivity/specificity for detection of pathogenic events.

CONCLUSION: Scripps Genome ADVISER CNV is designed to allow users with no prior

bioinformatics expertise to manipulate large volumes of copy-number variant data.

Scripps Genome ADVISER CNV is available at http://genomics.scripps.edu/ADVISER/.

DOI: 10.1038/gim.2014.180

PMCID: PMC4886732

PMID: 25521334 [Indexed for MEDLINE]

448. Guang Pu Xue Yu Guang Pu Fen Xi. 2015 Sep;35(9):2469-72.

[A Terahertz Spectral Database Based on Browser/Server Technique].

[Article in Chinese]

Zhang ZY, Song Y.

With the solution of key scientific and technical problems and development of

instrumentation, the application of terahertz technology in various fields has

been paid more and more attention. Owing to the unique characteristic advantages,

terahertz technology has been showing a broad future in the fields of fast,

non-damaging detections, as well as many other fields. Terahertz technology

combined with other complementary methods can be used to cope with many difficult

practical problems which could not be solved before. One of the critical points

for further development of practical terahertz detection methods depends on a

good and reliable terahertz spectral database. We developed a BS (browser/server)

-based terahertz spectral database recently. We designed the main structure and

main functions to fulfill practical requirements. The terahertz spectral database

now includes more than 240 items, and the spectral information was collected

based on three sources: (1) collection and citation from some other abroad

terahertz spectral databases; (2) collected from published literatures; and (3)

spectral data measured in our laboratory. The present paper introduced the basic

structure and fundament functions of the terahertz spectral database developed in

our laboratory. One of the key functions of this THz database is calculation of

optical parameters. Some optical parameters including absorption coefficient,

refractive index, etc. can be calculated based on the input THz time domain

spectra. The other main functions and searching methods of the

browser/server-based terahertz spectral database have been discussed. The

database search system can provide users convenient functions including user

registration, inquiry, displaying spectral figures and molecular structures,

spectral matching, etc. The THz database system provides an on-line searching

function for registered users. Registered users can compare the input THz

spectrum with the spectra of database, according to the obtained correlation

coefficient one can perform the searching task very fast and conveniently. Our

terahertz spectral database can be accessed at http://www.teralibrary.com. The

proposed terahertz spectral database is based on spectral information so far, and

will be improved in the future. We hope this terahertz spectral database can

provide users powerful, convenient, and high efficient functions, and could

promote the broader applications of terahertz technology.

PMID: 26669149

449. IEEE/ACM Trans Comput Biol Bioinform. 2015 Sep-Oct;12(5):1213-6. doi:

10.1109/TCBB.2015.2424411.

ZoomOut: Analyzing Multiple Networks as Single Nodes.

Athanasiadis EI, Bourdakou MM, Spyrou GM.

We have developed ZoomOut web server in order to provide the research community

with a tool for analysis, visualization and clustering of networks as a super

network, based on their calculated feature properties. Networks can be analysed

and be further treated as single nodes in a super network that describe their

relations. Specifically, the user interface is divided into three main sections:

the Workspace, the Networks Feature Calculations and the Clustering Networks

section. In the Workspace section, users are able to upload and manage multiple

networks for further processing. In the Networks Feature Calculations section, a

variety of network properties are calculated as features for each uploaded

network. In the Clustering Networks section, users are able to apply clustering

by selecting from the list of previously calculated features. All processed

networks can also be visualized as a super interactive network, were

interconnections among networks are based on the calculated clustering distances.

To the best of our knowledge, this is the first available web-service that allows

users to manage, quantify and visualize multiple networks at the same time,

handling them as parts of a larger network with properties calculated in an upper

scale. The ZoomOut web-application is available at

http://bioserver-3.bioacademy.gr/Bioserver/ZoomOut.

DOI: 10.1109/TCBB.2015.2424411

PMID: 26451833 [Indexed for MEDLINE]

450. Mol Biol (Mosk). 2015 Sep-Oct;49(5):846-53. doi: 10.7868/S0026898415050195.

[Web server for prediction of miRNAs and their precursors and binding sites].

[Article in Russian]

Vorozheykin PS(1), Titov II(1,)(2,)(3).

Author information:

(1)Novosibirsk State University, Novosibirsk, 630090, Russia. (2)Institute of

Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences,

Novosibirsk, 630090, Russia. (3)titov@bionet.nsc.ru.

A microRNA (miRNA) is a small noncoding RNA molecule about 22 nucleotides in

length. The paper describes a web server for predicting miRNAs and their

precursors and binding sites. The predictions are based on either sequence

similarity to known miRNAs of 223 organisms or context-structural hidden Markov

models. It has been shown that the proposed methods of prediction of human miRNAs

and pre-miRNAs outperform the existing ones in accuracy. The average deviation of

predicted 5'-ends of human miRNAs from actual positions is 3.13 nt in the case of

predicting one pair of complementary miRNAs (miRNA-miRNA\* duplex). A useful

option for our application is the prediction of an additional miRNA pair. In this

mode, the pairs closest to actual miRNA deviate by 1.61 nt on average. The

proposed method also shows good performance in predicting mouse miRNAs. Binding

sites for miRNAs are predicted by two known approaches based on complementarity

and thermodynamic stability of the miRNA-mRNA duplex and on a new approach, which

takes into account miRNAs competition for the site. The role of the secondary

structure in miRNA processing is considered. The web server is available at

http://wwwmgs.bionet.nsc.ru/mgs/programs/rnaanalys/.

DOI: 10.7868/S0026898415050195

PMID: 26510603 [Indexed for MEDLINE]

451. OMICS. 2015 Sep;19(9):574-7. doi: 10.1089/omi.2015.0060. Epub 2015 Aug 10.

LCGserver: A Webserver for Exploring Evolutionary Trajectory of Gene Orders in a

Large Number of Genomes.

Wang D(1,)(2), Yu J(1).

Author information:

(1)1 CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of

Genomics , Chinese Academy of Sciences, Beijing, People's Republic of China .

(2)2 Stem Cell Laboratory, UCL Cancer Institute, University College London ,

London, United Kingdom .

Genes and chromosomes are highly organized; together with protein-coding

sequence, gene structure at per gene level and gene order at cluster level are

both variable in a context of lineages and under natural selection. How gene

order and chromosome organization are related and selected remains to be

illuminated. The number of newly-sequenced genomes from various taxa has been

increasing rapidly, but there have not been easy-to-use web tools that allow

better visualization for gene order in a large genome collection. Here, we

describe a webserver, LCGserver (http://lcgbase.big.ac.cn/LCGserver/), for

exploring evolutionary dynamics of gene orders over diverse lineages. This server

provides gene order information at three levels: single gene, paired gene (a

minimal cluster), and clustered gene (more than two genes). The most exclusive

feature of LCGserver is alignment and visualization of neighboring genes based on

orthology, allowing users to inspect all conserved and dynamic events of gene

order along chromosomes in a lineage-specific manner. In addition, it categories

paired genes into six patterns and identifies fully-conserved gene clusters

within and among lineages.

DOI: 10.1089/omi.2015.0060

PMCID: PMC4575520

PMID: 26258441 [Indexed for MEDLINE]

452. Physiol Plant. 2015 Sep;155(1):12-20. doi: 10.1111/ppl.12326. Epub 2015 Feb 24.

The MORPH-R web server and software tool for predicting missing genes in

biological pathways.

Amar D(1), Frades I(2), Diels T(3,)(4), Zaltzman D(5), Ghatan N(5), Hedley PE(6),

Alexandersson E(2), Tzfadia O(3,)(5), Shamir R(1).

Author information:

(1)Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel.

(2)Department of Plant Protection Biology, Swedish University of Agricultural

Sciences, Alnarp, Sweden. (3)Department of Plant Systems Biology, VIB, Ghent,

Belgium. (4)Department of Mathematics and Computer Science, University of

Antwerp, Antwerp, Belgium. (5)Department of Plant Science, The Weizmann Institute

of Science, Rehovot, Israel. (6)Cell and Molecular Sciences, The James Hutton

Institute, Dundee, United Kingdom.

A biological pathway is the set of molecular entities involved in a given

biological process and the interrelations among them. Even though biological

pathways have been studied extensively, discovering missing genes in pathways

remains a fundamental challenge. Here, we present an easy-to-use tool that allows

users to run MORPH (MOdule-guided Ranking of candidate PatHway genes), an

algorithm for revealing missing genes in biological pathways, and demonstrate its

capabilities. MORPH supports the analysis in tomato, Arabidopsis and the two new

species: rice and the newly sequenced potato genome. The new tool, called

MORPH-R, is available both as a web server (at

http://bioinformatics.psb.ugent.be/webtools/morph/) and as standalone software

that can be used locally. In the standalone version, the user can apply the tool

to new organisms using any proprietary and public data sources.

© 2015 Scandinavian Plant Physiology Society.

DOI: 10.1111/ppl.12326

PMID: 25625434 [Indexed for MEDLINE]

453. Bioinformation. 2015 Aug 31;11(8):416-21. doi: 10.6026/97320630011416.

eCollection 2015.

NNvPDB: Neural Network based Protein Secondary Structure Prediction with PDB

Validation.

Sakthivel S(1), S K M H(1).

Author information:

(1)Department of Bioinformatics, School of Bioengineering, Faculty of Engineering

& Technology, Kattankulathur Campus, SRM University, Potheri - 603203, Tamil

Nadu, India.

The predicted secondary structural states are not cross validated by any of the

existing servers. Hence, information on the level of accuracy for every sequence

is not reported by the existing servers. This was overcome by NNvPDB, which not

only reported greater Q3 but also validates every prediction with the homologous

PDB entries. NNvPDB is based on the concept of Neural Network, with a new and

different approach of training the network every time with five PDB structures

that are similar to query sequence. The average accuracy for helix is 76%, beta

sheet is 71% and overall (helix, sheet and coil) is 66%.AVAILABILITY:

http://bit.srmuniv.ac.in/cgi-bin/bit/cfpdb/nnsecstruct.pl.

DOI: 10.6026/97320630011416

PMCID: PMC4574126

PMID: 26420924

454. PLoS One. 2015 Aug 28;10(8):e0136711. doi: 10.1371/journal.pone.0136711.

eCollection 2015.

FMiR: A Curated Resource of Mitochondrial DNA Information for Fish.

Nagpure NS(1), Rashid I(1), Pathak AK(2), Singh M(1), Pati R(1), Singh SP(2),

Sarkar UK(2).

Author information:

(1)Division of Molecular Biology and Biotechnology, National Bureau of Fish

Genetic Resources, Lucknow-226002, India. (2)Division of Fish Taxonomy and

Resources, National Bureau of Fish Genetic Resources, Lucknow-226002, India.

Mitochondrial genome sequences have been widely used for evolutionary and

phylogenetic studies. Among vertebrates, fish are an important, diverse group,

and their mitogenome sequences are growing rapidly in public repositories. To

facilitate mitochondrial genome analysis and to explore the valuable genetic

information, we developed the Fish Mitogenome Resource (FMiR) database to provide

a workbench for mitogenome annotation, species identification and microsatellite

marker mining. The microsatellites are also known as simple sequence repeats

(SSRs) and used as molecular markers in studies on population genetics, gene

duplication and marker assisted selection. Here, easy-to-use tools have been

implemented for mining SSRs and for designing primers to identify species/habitat

specific markers. In addition, FMiR can analyze complete or partial mitochondrial

genome sequence to identify species and to deduce relational distances among

sequences across species. The database presently contains curated mitochondrial

genomes from 1302 fish species belonging to 297 families and 47 orders reported

from saltwater and freshwater ecosystems. In addition, the database covers

information on fish species such as conservation status, ecosystem, family,

distribution and occurrence downloaded from the FishBase and IUCN Red List

databases. Those fish information have been used to browse mitogenome information

for the species belonging to a particular category. The database is scalable in

terms of content and inclusion of other analytical modules. The FMiR is running

under Linux operating platform on high performance server accessible at URL

http://mail.nbfgr.res.in/fmir.

DOI: 10.1371/journal.pone.0136711

PMCID: PMC4552752

PMID: 26317619 [Indexed for MEDLINE]

455. J Phys Chem B. 2015 Aug 27;119(34):11136-45. doi: 10.1021/acs.jpcb.5b02999. Epub

2015 Jul 2.

From Ramachandran Maps to Tertiary Structures of Proteins.

DasGupta D(1), Kaushik R(1), Jayaram B(1).

Author information:

(1)Department of Chemistry, ‡Supercomputing Facility for Bioinformatics &

Computational Biology, and §Kusuma School of Biological Sciences, Indian

Institute of Technology , Hauz Khas, New Delhi-110016, India.

Sequence to structure of proteins is an unsolved problem. A possible coarse

grained resolution to this entails specification of all the torsional (Φ, Ψ)

angles along the backbone of the polypeptide chain. The Ramachandran map quite

elegantly depicts the allowed conformational (Φ, Ψ) space of proteins which is

still very large for the purposes of accurate structure generation. We have

divided the allowed (Φ, Ψ) space in Ramachandran maps into 27 distinct

conformations sufficient to regenerate a structure to within 5 Å from the native,

at least for small proteins, thus reducing the structure prediction problem to a

specification of an alphanumeric string, i.e., the amino acid sequence together

with one of the 27 conformations preferred by each amino acid residue. This still

theoretically results in 27(n) conformations for a protein comprising "n" amino

acids. We then investigated the spatial correlations at the two-residue

(dipeptide) and three-residue (tripeptide) levels in what may be described as

higher order Ramachandran maps, with the premise that the allowed conformational

space starts to shrink as we introduce neighborhood effects. We found, for

instance, for a tripeptide which potentially can exist in any of the 27(3)

"allowed" conformations, three-fourths of these conformations are redundant to

the 95% confidence level, suggesting sequence context dependent preferred

conformations. We then created a look-up table of preferred conformations at the

tripeptide level and correlated them with energetically favorable conformations.

We found in particular that Boltzmann probabilities calculated from van der Waals

energies for each conformation of tripeptides correlate well with the observed

populations in the structural database (the average correlation coefficient is

∼0.8). An alpha-numeric string and hence the tertiary structure can be generated

for any sequence from the look-up table within minutes on a single processor and

to a higher level of accuracy if secondary structure can be specified. We tested

the methodology on 100 small proteins, and in 90% of the cases, a structure

within 5 Å is recovered. We thus believe that the method presented here provides

the missing link between Ramachandran maps and tertiary structures of proteins. A

Web server to convert a tertiary structure to an alphanumeric string and to

predict the tertiary structure from the sequence of a protein using the above

methodology is created and made freely accessible at

http://www.scfbio-iitd.res.in/software/proteomics/rm2ts.jsp.

DOI: 10.1021/acs.jpcb.5b02999

PMID: 26098815 [Indexed for MEDLINE]

456. J Theor Biol. 2015 Aug 21;379:10-5. doi: 10.1016/j.jtbi.2015.04.016. Epub 2015

Apr 24.

Phogly-PseAAC: Prediction of lysine phosphoglycerylation in proteins

incorporating with position-specific propensity.

Xu Y(1), Ding YX(2), Ding J(2), Wu LY(3), Deng NY(4).

Author information:

(1)Department of Information and Computer Science, University of Science and

Technology, Beijing, China. Electronic address: xuyan@ustb.edu.cn. (2)Department

of Information and Computer Science, University of Science and Technology,

Beijing, China. (3)Institute of Applied Mathematics, Academy of Mathematics and

Systems Science, Chinese Academy of Sciences, Beijing 100190, China. (4)College

of Science, China Agricultural University, Beijing 100083, China.

Large-scale characterization of post-translational modifications (PTMs), such as

posphorylation, acetylation and ubiquitination, has highlighted their importance

in the regulation of a myriad of signaling events. However, as another type of

PTMs-lysine phosphoglycerylation, the data of phosphoglycerylated sites has just

been manually experimented in recent years. Given an uncharacterized protein

sequence that contains many lysine residues, which one of them can be

phosphoglycerylated and which one not? This is a challenging problem. In view of

this, establishing a useful computational method and developing an efficient

predictor are highly desired. Here a new predictor named Phogly-PseAAC was

developed which incorporated with the position specific amino acid propensity.

The feature importance through F-score value has also been ranked. The predictor

with the best feature set obtained the accuracy 75.10%, sensitivity 68.87%,

specificity 75.57% and MCC 0.2538 in LOO test cross validation with center

nearest neighbor algorithm. Meanwhile, a web-server for Phogly-PseAAC is

accessible at http://app.aporc.org/Phogly-PseAAC/. For the convenience of most

experimental scientists, we have further provided a brief instruction for the

web-server, by which users can easily get their desired results without the need

to follow the complicated mathematics presented in this paper. It is anticipated

that Phogly-PseAAC may become a useful high throughput tool for identifying the

lysine phosphoglycerylation sites.

Copyright © 2015 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2015.04.016

PMID: 25913879 [Indexed for MEDLINE]

457. Bioinformatics. 2015 Aug 15;31(16):2639-45. doi: 10.1093/bioinformatics/btv212.

Epub 2015 Apr 20.

MultiP-SChlo: multi-label protein subchloroplast localization prediction with

Chou's pseudo amino acid composition and a novel multi-label classifier.

Wang X(1), Zhang W(1), Zhang Q(1), Li GZ(2).

Author information:

(1)School of Computer and Communication Engineering, Zhengzhou University of

Light Industry, Zhengzhou 450002, China and. (2)Department of Control Science and

Engineering, Tongji University, Shanghai 201804, China.

MOTIVATION: Identifying protein subchloroplast localization in chloroplast

organelle is very helpful for understanding the function of chloroplast proteins.

There have existed a few computational prediction methods for protein

subchloroplast localization. However, these existing works have ignored proteins

with multiple subchloroplast locations when constructing prediction models, so

that they can predict only one of all subchloroplast locations of this kind of

multilabel proteins.

RESULTS: To address this problem, through utilizing label-specific features and

label correlations simultaneously, a novel multilabel classifier was developed

for predicting protein subchloroplast location(s) with both single and multiple

location sites. As an initial study, the overall accuracy of our proposed

algorithm reaches 55.52%, which is quite high to be able to become a promising

tool for further studies.

AVAILABILITY AND IMPLEMENTATION: An online web server for our proposed algorithm

named MultiP-SChlo was developed, which are freely accessible at

http://biomed.zzuli.edu.cn/bioinfo/multip-schlo/.

CONTACT: pandaxiaoxi@gmail.com or gzli@tongji.edu.cn

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv212

PMID: 25900916 [Indexed for MEDLINE]

458. Bioinformatics. 2015 Aug 15;31(16):2748-50. doi: 10.1093/bioinformatics/btv200.

Epub 2015 Apr 9.

AVIA v2.0: annotation, visualization and impact analysis of genomic variants and

genes.

Vuong H(1), Che A(1), Ravichandran S(1), Luke BT(1), Collins JR(1), Mudunuri

US(1).

Author information:

(1)Advanced Biomedical Computing Center, Frederick National Laboratory for Cancer

Research, Frederick, MD 21702, USA.

As sequencing becomes cheaper and more widely available, there is a greater need

to quickly and effectively analyze large-scale genomic data. While the

functionality of AVIA v1.0, whose implementation was based on ANNOVAR, was

comparable with other annotation web servers, AVIA v2.0 represents an enhanced

web-based server that extends genomic annotations to cell-specific transcripts

and protein-level functional annotations. With AVIA's improved interface, users

can better visualize their data, perform comprehensive searches and categorize

both coding and non-coding variants.AVAILABILITY AND IMPLEMENTATION: AVIA is

freely available through the web at http://avia.abcc.ncifcrf.gov.

CONTACT: Hue.Vuong@fnlcr.nih.gov

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv200

PMCID: PMC4528632

PMID: 25861966 [Indexed for MEDLINE]

459. Bioinformatics. 2015 Aug 15;31(16):2745-7. doi: 10.1093/bioinformatics/btv195.

Epub 2015 Apr 6.

PROVEAN web server: a tool to predict the functional effect of amino acid

substitutions and indels.

Choi Y(1), Chan AP(1).

Author information:

(1)The J. Craig Venter Institute, Rockville, MD 20850, USA.

We present a web server to predict the functional effect of single or multiple

amino acid substitutions, insertions and deletions using the prediction tool

PROVEAN. The server provides rapid analysis of protein variants from any

organisms, and also supports high-throughput analysis for human and mouse

variants at both the genomic and protein levels.AVAILABILITY AND IMPLEMENTATION:

The web server is freely available and open to all users with no login

requirements at http://provean.jcvi.org.

CONTACT: achan@jcvi.org

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv195

PMCID: PMC4528627

PMID: 25851949 [Indexed for MEDLINE]

460. Database (Oxford). 2015 Aug 8;2015. pii: bav076. doi: 10.1093/database/bav076.

Print 2015.

MESSI: metabolic engineering target selection and best strain identification

tool.

Kang K(1), Li J(1), Lim BL(2), Panagiotou G(3).

Author information:

(1)Systems Biology & Bioinformatics Group, School of Biological Sciences, The

University of Hong Kong, Pokfulam Road, Hong Kong and. (2)School of Biological

Sciences, The University of Hong Kong, Pokfulam Road, Hong Kong. (3)Systems

Biology & Bioinformatics Group, School of Biological Sciences, The University of

Hong Kong, Pokfulam Road, Hong Kong and gipa@hku.hk.

Metabolic engineering and synthetic biology are synergistically related fields

for manipulating target pathways and designing microorganisms that can act as

chemical factories. Saccharomyces cerevisiae's ideal bioprocessing traits make

yeast a very attractive chemical factory for production of fuels,

pharmaceuticals, nutraceuticals as well as a wide range of chemicals. However,

future attempts of engineering S. cerevisiae's metabolism using synthetic biology

need to move towards more integrative models that incorporate the high

connectivity of metabolic pathways and regulatory processes and the interactions

in genetic elements across those pathways and processes. To contribute in this

direction, we have developed Metabolic Engineering target Selection and best

Strain Identification tool (MESSI), a web server for predicting efficient chassis

and regulatory components for yeast bio-based production. The server provides an

integrative platform for users to analyse ready-to-use public high-throughput

metabolomic data, which are transformed to metabolic pathway activities for

identifying the most efficient S. cerevisiae strain for the production of a

compound of interest. As input MESSI accepts metabolite KEGG IDs or pathway

names. MESSI outputs a ranked list of S. cerevisiae strains based on aggregation

algorithms. Furthermore, through a genome-wide association study of the metabolic

pathway activities with the strains' natural variation, MESSI prioritizes genes

and small variants as potential regulatory points and promising metabolic

engineering targets. Users can choose various parameters in the whole process

such as (i) weight and expectation of each metabolic pathway activity in the

final ranking of the strains, (ii) Weighted AddScore Fuse or Weighted Borda Fuse

aggregation algorithm, (iii) type of variants to be included, (iv) variant sets

in different biological levels.Database URL: http://sbb.hku.hk/MESSI/.

© The Author(s) 2015. Published by Oxford University Press.

DOI: 10.1093/database/bav076

PMCID: PMC4529744

PMID: 26255308 [Indexed for MEDLINE]

461. Biol Direct. 2015 Aug 5;10:41. doi: 10.1186/s13062-015-0070-9.

A membrane computing simulator of trans-hierarchical antibiotic resistance

evolution dynamics in nested ecological compartments (ARES).

Campos M(1,)(2), Llorens C(3), Sempere JM(4), Futami R(5), Rodriguez

I(6,)(7,)(8), Carrasco P(9), Capilla R(10), Latorre A(11,)(12,)(13), Coque

TM(14,)(15,)(16), Moya A(17,)(18,)(19), Baquero F(20,)(21,)(22).

Author information:

(1)Department of Microbiology, Ramón y Cajal University Hospital, IRYCIS,

Carretera de Colmenar Viejo, km. 9,100, 28034, Madrid, Spain.

mcampos@dsic.upv.es. (2)Department of Information Systems and Computation (DSIC),

Polytechnic University of Valencia, Camino de Vera, 46022, Valencia, Spain.

mcampos@dsic.upv.es. (3)Biotechvana, Valencia, CEEI Building, Benjamin Franklin

Av. 12, Valencia Technological Park, 46980, Paterna, Spain.

carlos.llorens@biotechvana.com. (4)Department of Information Systems and

Computation (DSIC), Polytechnic University of Valencia, Camino de Vera, 46022,

Valencia, Spain. jsempere@dsic.upv.es. (5)Biotechvana, Valencia, CEEI Building,

Benjamin Franklin Av. 12, Valencia Technological Park, 46980, Paterna, Spain.

ricardo.futami@biotechvana.com. (6)Department of Microbiology, Ramón y Cajal

University Hospital, IRYCIS, Carretera de Colmenar Viejo, km. 9,100, 28034,

Madrid, Spain. irene.rodriguez.fdez@gmail.com. (7)Antibiotic Resistance and

Bacterial Virulence Unit (HRYC-CSIC), associated to the Superior Council of

Scientific Investigations (CSIC), Madrid, Spain. irene.rodriguez.fdez@gmail.com.

(8)Network Research Center for Epidemiology and Public Health (CIBER-ESP),

Madrid, Spain. irene.rodriguez.fdez@gmail.com. (9)Cavanilles Institute of

Biodiversity and Evolutionary Biology, University of Valencia, C/ Catedrático

José Beltrán 2, 46980, Paterna, Valencia, Spain. purificacion.carrasco@uv.es.

(10)Biotechvana, Valencia, CEEI Building, Benjamin Franklin Av. 12, Valencia

Technological Park, 46980, Paterna, Spain. rafael.capilla@biotechvana.com.

(11)Network Research Center for Epidemiology and Public Health (CIBER-ESP),

Madrid, Spain. amparo.latorre@uv.es. (12)Cavanilles Institute of Biodiversity and

Evolutionary Biology, University of Valencia, C/ Catedrático José Beltrán 2,

46980, Paterna, Valencia, Spain. amparo.latorre@uv.es. (13)Foundation for the

Promotion of Health and Biomedical Research in the Valencian Community (FISABIO)

- Public Health, Avenida de Cataluña 21, 46020, Valencia, Spain.

amparo.latorre@uv.es. (14)Department of Microbiology, Ramón y Cajal University

Hospital, IRYCIS, Carretera de Colmenar Viejo, km. 9,100, 28034, Madrid, Spain.

teresacoque@gmail.com. (15)Antibiotic Resistance and Bacterial Virulence Unit

(HRYC-CSIC), associated to the Superior Council of Scientific Investigations

(CSIC), Madrid, Spain. teresacoque@gmail.com. (16)Network Research Center for

Epidemiology and Public Health (CIBER-ESP), Madrid, Spain. teresacoque@gmail.com.

(17)Network Research Center for Epidemiology and Public Health (CIBER-ESP),

Madrid, Spain. andres.moya@uv.es. (18)Cavanilles Institute of Biodiversity and

Evolutionary Biology, University of Valencia, C/ Catedrático José Beltrán 2,

46980, Paterna, Valencia, Spain. andres.moya@uv.es. (19)Foundation for the

Promotion of Health and Biomedical Research in the Valencian Community (FISABIO)

- Public Health, Avenida de Cataluña 21, 46020, Valencia, Spain.

andres.moya@uv.es. (20)Department of Microbiology, Ramón y Cajal University

Hospital, IRYCIS, Carretera de Colmenar Viejo, km. 9,100, 28034, Madrid, Spain.

baquero@bitmailer.net. (21)Antibiotic Resistance and Bacterial Virulence Unit

(HRYC-CSIC), associated to the Superior Council of Scientific Investigations

(CSIC), Madrid, Spain. baquero@bitmailer.net. (22)Network Research Center for

Epidemiology and Public Health (CIBER-ESP), Madrid, Spain. baquero@bitmailer.net.

BACKGROUND: Antibiotic resistance is a major biomedical problem upon which public

health systems demand solutions to construe the dynamics and epidemiological risk

of resistant bacteria in anthropogenically-altered environments. The

implementation of computable models with reciprocity within and between levels of

biological organization (i.e. essential nesting) is central for studying

antibiotic resistances. Antibiotic resistance is not just the result of

antibiotic-driven selection but more properly the consequence of a complex

hierarchy of processes shaping the ecology and evolution of the distinct

subcellular, cellular and supra-cellular vehicles involved in the dissemination

of resistance genes. Such a complex background motivated us to explore the

P-system standards of membrane computing an innovative natural computing

formalism that abstracts the notion of movement across membranes to simulate

antibiotic resistance evolution processes across nested levels of micro- and

macro-environmental organization in a given ecosystem.

RESULTS: In this article, we introduce ARES (Antibiotic Resistance Evolution

Simulator) a software device that simulates P-system model scenarios with five

types of nested computing membranes oriented to emulate a hierarchy of

eco-biological compartments, i.e. a) peripheral ecosystem; b) local environment;

c) reservoir of supplies; d) animal host; and e) host's associated bacterial

organisms (microbiome). Computational objects emulating molecular entities such

as plasmids, antibiotic resistance genes, antimicrobials, and/or other substances

can be introduced into this framework and may interact and evolve together with

the membranes, according to a set of pre-established rules and specifications.

ARES has been implemented as an online server and offers additional tools for

storage and model editing and downstream analysis.

CONCLUSIONS: The stochastic nature of the P-system model implemented in ARES

explicitly links within and between host dynamics into a simulation, with

feedback reciprocity among the different units of selection influenced by

antibiotic exposure at various ecological levels. ARES offers the possibility of

modeling predictive multilevel scenarios of antibiotic resistance evolution that

can be interrogated, edited and re-simulated if necessary, with different

parameters, until a correct model description of the process in the real world is

convincingly approached. ARES can be accessed at http://gydb.org/ares.

DOI: 10.1186/s13062-015-0070-9

PMCID: PMC4526193

PMID: 26243297 [Indexed for MEDLINE]

462. F1000Res. 2015 Aug 5;4:477. doi: 10.12688/f1000research.6773.1. eCollection 2015.

SLiMScape 3.x: a Cytoscape 3 app for discovery of Short Linear Motifs in protein

interaction networks.

Olorin E(1), O'Brien KT(2), Palopoli N(3), Pérez-Bercoff Å(1), Shields DC(2),

Edwards RJ(4).

Author information:

(1)School of Biotechnology and Biomolecular Sciences, University of New South

Wales, Sydney, Australia. (2)UCD Conway Institute of Biomolecular and Biomedical

Research, School of Medicine, University College Dublin, Dublin, Ireland.

(3)Centre for Biological Sciences, University of Southampton, Southampton, UK ;

Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal,

Argentina ; Fundación Instituto Leloir, Buenos Aires, Argentina. (4)School of

Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney,

Australia ; Centre for Biological Sciences, University of Southampton,

Southampton, UK.

Short linear motifs (SLiMs) are small protein sequence patterns that mediate a

large number of critical protein-protein interactions, involved in processes

such as complex formation, signal transduction, localisation and stabilisation.

SLiMs show rapid evolutionary dynamics and are frequently the targets of

molecular mimicry by pathogens. Identifying enriched sequence patterns due to

convergent evolution in non-homologous proteins has proven to be a successful

strategy for computational SLiM prediction. Tools of the SLiMSuite package use

this strategy, using a statistical model to identify SLiM enrichment based on the

evolutionary relationships, amino acid composition and predicted disorder of the

input proteins. The quality of input data is critical for successful SLiM

prediction. Cytoscape provides a user-friendly, interactive environment to

explore interaction networks and select proteins based on common features, such

as shared interaction partners. SLiMScape embeds tools of the SLiMSuite package

for de novo SLiM discovery (SLiMFinder and QSLiMFinder) and identifying

occurrences/enrichment of known SLiMs (SLiMProb) within this interactive

framework. SLiMScape makes it easier to (1) generate high quality

hypothesis-driven datasets for these tools, and (2) visualise predicted SLiM

occurrences within the context of the network. To generate new predictions, users

can select nodes from a protein network or provide a set of Uniprot identifiers.

SLiMProb also requires additional query motif input. Jobs are then run remotely

on the SLiMSuite server ( http://rest.slimsuite.unsw.edu.au) for subsequent

retrieval and visualisation. SLiMScape can also be used to retrieve and visualise

results from jobs run directly on the server. SLiMScape and SLiMSuite are open

source and freely available via GitHub under GNU licenses.

DOI: 10.12688/f1000research.6773.1

PMCID: PMC4670012

PMID: 26674271

463. Biosystems. 2015 Aug;134:37-42. doi: 10.1016/j.biosystems.2015.05.004. Epub 2015

Jun 17.

miSEA: microRNA set enrichment analysis.

Çorapçıoğlu ME(1), Oğul H(2).

Author information:

(1)Department of Computer Engineering, Baskent University, Ankara, Turkey;

Department of Technology and Information Management, Baskent University, Ankara,

Turkey. (2)Department of Computer Engineering, Baskent University, Ankara,

Turkey. Electronic address: hogul@baskent.edu.tr.

We introduce a novel web-based tool, miSEA, for evaluating the enrichment of

relevant microRNA sets from microarray and miRNA-Seq experiments on paired

samples, e.g. control vs.TREATMENT: In addition to a group of previously

annotated microRNA sets embedded in the system, this tool enables users to import

new microRNA sets obtained from their own research. miSEA allows users to select

from a large variety of microRNA grouping categories, such as family

classification, disease association, common regulation, and genome coordinates,

based on their requirements. miSEA therefore provides a knowledge-driven

representation scheme for microRNA experiments. The usability of this platform

was discerned with a cancer type-classification task performed on a set of real

microRNA expression profiling experiments. The miSEA web server is available at

http://www.baskent.edu.tr/∼hogul/misea.

Copyright © 2015 Elsevier Ireland Ltd. All rights reserved.

DOI: 10.1016/j.biosystems.2015.05.004

PMID: 26093049 [Indexed for MEDLINE]

464. J Membr Biol. 2015 Aug;248(4):745-52. doi: 10.1007/s00232-015-9787-8. Epub 2015

Mar 22.

iMem-Seq: A Multi-label Learning Classifier for Predicting Membrane Proteins

Types.

Xiao X(1), Zou HL, Lin WZ.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen, 333046,

China, jdzxiaoxuan@163.com.

Predicting membrane protein type is a challenging problem, particularly when the

query proteins may simultaneously have two or more different types. Most of the

existing methods can only be used to deal with the single-label proteins.

Actually, multiple-label proteins should not be ignored because they usually bear

some special functions worthy of in-depth studies. By introducing the

"multi-labeled learning" and hybridizing evolution information through Grey-PSSM,

a novel predictor called iMem-Seq is developed that can be used to deal with the

systems containing both single and multiple types of membrane proteins. As a

demonstration, the jackknife cross-validation was performed with iMem-Seq on a

benchmark dataset of membrane proteins classified into the eight types, where

some proteins belong to two or there types, but none has ≥25% pairwise sequence

identity to any other in a same subset. It was demonstrated via the rigorous

cross-validations that the new predictor remarkably outperformed all its

counterparts. As a user-friendly web-server, iMem-Seq is freely accessible to the

public at the website http://www.jci-bioinfo.cn/iMem-Seq .

DOI: 10.1007/s00232-015-9787-8

PMID: 25796484 [Indexed for MEDLINE]

465. Proteins. 2015 Aug;83(8):1436-49. doi: 10.1002/prot.24829. Epub 2015 Jun 6.

CONFOLD: Residue-residue contact-guided ab initio protein folding.

Adhikari B(1), Bhattacharya D(1), Cao R(1), Cheng J(1).

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, Missouri,

65211.

Predicted protein residue-residue contacts can be used to build three-dimensional

models and consequently to predict protein folds from scratch. A considerable

amount of effort is currently being spent to improve contact prediction accuracy,

whereas few methods are available to construct protein tertiary structures from

predicted contacts. Here, we present an ab initio protein folding method to build

three-dimensional models using predicted contacts and secondary structures. Our

method first translates contacts and secondary structures into distance, dihedral

angle, and hydrogen bond restraints according to a set of new conversion rules,

and then provides these restraints as input for a distance geometry algorithm to

build tertiary structure models. The initially reconstructed models are used to

regenerate a set of physically realistic contact restraints and detect secondary

structure patterns, which are then used to reconstruct final structural models.

This unique two-stage modeling approach of integrating contacts and secondary

structures improves the quality and accuracy of structural models and in

particular generates better β-sheets than other algorithms. We validate our

method on two standard benchmark datasets using true contacts and secondary

structures. Our method improves TM-score of reconstructed protein models by 45%

and 42% over the existing method on the two datasets, respectively. On the

dataset for benchmarking reconstructions methods with predicted contacts and

secondary structures, the average TM-score of best models reconstructed by our

method is 0.59, 5.5% higher than the existing method. The CONFOLD web server is

available at http://protein.rnet.missouri.edu/confold/.

© 2015 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24829

PMCID: PMC4509844

PMID: 25974172 [Indexed for MEDLINE]

466. PLoS One. 2015 Jul 24;10(7):e0133647. doi: 10.1371/journal.pone.0133647.

eCollection 2015.

GREAM: A Web Server to Short-List Potentially Important Genomic Repeat Elements

Based on Over-/Under-Representation in Specific Chromosomal Locations, Such as

the Gene Neighborhoods, within or across 17 Mammalian Species.

Chandrashekar DS(1), Dey P(1), Acharya KK(2).

Author information:

(1)Institute of Bioinformatics and Applied Biotechnology (IBAB), Biotech Park,

Electronic City, Bengaluru (Bangalore), 560100, Karnataka state, India; Manipal

University, Manipal, 576104, Karnataka state, India. (2)Institute of

Bioinformatics and Applied Biotechnology (IBAB), Biotech Park, Electronic City,

Bengaluru (Bangalore), 560100, Karnataka state, India; Shodhaka Life Sciences

Pvt. Ltd., IBAB, Biotech Park, Bengaluru (Bangalore), 560100, Karnataka state,

India.

BACKGROUND: Genome-wide repeat sequences, such as LINEs, SINEs and LTRs share a

considerable part of the mammalian nuclear genomes. These repeat elements seem to

be important for multiple functions including the regulation of transcription

initiation, alternative splicing and DNA methylation. But it is not possible to

study all repeats and, hence, it would help to short-list before exploring their

potential functional significance via experimental studies and/or detailed in

silico analyses.

RESULT: We developed the 'Genomic Repeat Element Analyzer for Mammals' (GREAM)

for analysis, screening and selection of potentially important mammalian genomic

repeats. This web-server offers many novel utilities. For example, this is the

only tool that can reveal a categorized list of specific types of transposons,

retro-transposons and other genome-wide repetitive elements that are

statistically over-/under-represented in regions around a set of genes, such as

those expressed differentially in a disease condition. The output displays the

position and frequency of identified elements within the specified regions. In

addition, GREAM offers two other types of analyses of genomic repeat sequences:

a) enrichment within chromosomal region(s) of interest, and b) comparative

distribution across the neighborhood of orthologous genes. GREAM successfully

short-listed a repeat element (MER20) known to contain functional motifs. In

other case studies, we could use GREAM to short-list repetitive elements in the

azoospermia factor a (AZFa) region of the human Y chromosome and those around the

genes associated with rat liver injury. GREAM could also identify five

over-represented repeats around some of the human and mouse transcription factor

coding genes that had conserved expression patterns across the two species.

CONCLUSION: GREAM has been developed to provide an impetus to research on the

role of repetitive sequences in mammalian genomes by offering easy selection of

more interesting repeats in various contexts/regions. GREAM is freely available

at http://resource.ibab.ac.in/GREAM/.

DOI: 10.1371/journal.pone.0133647

PMCID: PMC4514817

PMID: 26208093 [Indexed for MEDLINE]

467. J Theor Biol. 2015 Jul 21;377:47-56. doi: 10.1016/j.jtbi.2015.04.011. Epub 2015

Apr 20.

iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by

incorporating their physicochemical properties and wavelet transforms into

PseAAC.

Jia J(1), Liu Z(2), Xiao X(3), Liu B(4), Chou KC(5).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China; School of Computer Science, University of Birmingham, Edgbaston Birmingham

B15 2TT, UK. Electronic address: jjh163yx@163.com. (2)Computer Department,

Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China. Electronic address:

liuzi189836@163.com. (3)Computer Department, Jing-De-Zhen Ceramic Institute,

Jing-De-Zhen 333403, China; Gordon Life Science Institute, Boston, MA 02478, USA.

Electronic address: xxiao@gordonlifescience.org. (4)Computer Department,

Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China. Electronic address:

lbx1966@163.com. (5)Center of Excellence in Genomic Medicine Research (CEGMR),

King Abdulaziz University, Jeddah 21589, Saudi Arabia; Gordon Life Science

Institute, Boston, MA 02478, USA. Electronic address:

kcchou@gordonlifescience.org.

A cell contains thousands of proteins. Many important functions of cell are

carried out through the proteins therein. Proteins rarely function alone. Most of

their functions essential to life are associated with various types of

protein-protein interactions (PPIs). Therefore, knowledge of PPIs is fundamental

for both basic research and drug development. With the avalanche of proteins

sequences generated in the postgenomic age, it is highly desired to develop

computational methods for timely acquiring this kind of knowledge. Here, a new

predictor, called "iPPI-Emsl", is developed. In the predictor, a protein sample

is formulated by incorporating the following two types of information into the

general form of PseAAC (pseudo amino acid composition): (1) the physicochemical

properties derived from the constituent amino acids of a protein; and (2) the

wavelet transforms derived from the numerical series along a protein chain. The

operation engine to run the predictor is an ensemble classifier formed by fusing

seven individual random forest engines via a voting system. It is demonstrated

with the benchmark dataset from Saccharomyces cerevisiae as well as the dataset

from Helicobacter pylori that the new predictor achieves remarkably higher

success rates than any of the existing predictors in this area. The new

predictor׳ web-server has been established at

http://www.jci-bioinfo.cn/iPPI-Esml. For the convenience of most experimental

scientists, we have further provided a step-by-step guide, by which users can

easily get their desired results without the need to follow the complicated

mathematics involved during its development.

Copyright © 2015 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2015.04.011

PMID: 25908206 [Indexed for MEDLINE]

468. Bioinformatics. 2015 Jul 15;31(14):2403-5. doi: 10.1093/bioinformatics/btv140.

Epub 2015 Mar 10.

miR2GO: comparative functional analysis for microRNAs.

Bhattacharya A(1), Cui Y(1).

Author information:

(1)Department of Microbiology, Immunology and Biochemistry and Center for

Integrative and Translational Genomics, University of Tennessee Health Science

Center, Memphis, TN 38163, USA.

miR2GO is a web-based platform for comparative analyses of human miRNA functions.

It includes two programs: miRmut2GO and miRpair2GO. miRmut2GO implements a

knowledge-based method to assess the functional effects of genetic and somatic

mutations in microRNA seed regions. The functional effects of a mutation are

analysed by semantic comparison of enriched gene ontology (GO) annotations of the

target gene sets for the wild-type and mutated alleles. miRpair2GO compares the

functions of two different miRNAs based on the enriched functional annotations of

their target gene sets.AVAILABILITY AND IMPLEMENTATION: The miR2GO web server is

available at http://compbio.uthsc.edu/miR2GO.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv140

PMID: 25762653 [Indexed for MEDLINE]

469. Bioinformatics. 2015 Jul 15;31(14):2388-90. doi: 10.1093/bioinformatics/btv136.

Epub 2015 Mar 6.

EpIC: a rational pipeline for epitope immunogenicity characterization.

Marciniuk K(1), Trost B(2), Napper S(1).

Author information:

(1)Department of Biochemistry, University of Saskatchewan, Saskatoon, SK, S7N 5E5

Canada, Vaccine and Infectious Disease Organization-International Vaccine Centre,

University of Saskatchewan, Saskatoon, SK, S7N 5E3 Canada and. (2)Vaccine and

Infectious Disease Organization-International Vaccine Centre, University of

Saskatchewan, Saskatoon, SK, S7N 5E3 Canada and Department of Computer Science,

University of Saskatchewan, Saskatoon, SK, S7N 5C9 Canada.

Efforts to develop peptide-based vaccines, in particular those requiring

site-specific targeting of self-proteins, rely on the ability to optimize the

immunogenicity of the peptide epitopes. Currently, screening of candidate

vaccines is typically performed through low-throughput, high-cost animal trials.

To improve on this we present the program EpIC, which enables high-throughput

prediction of peptide immunogenicity based on the endogenous occurrence of B-cell

epitopes within native protein sequences. This information informs rational

selection of immunogenicity-optimized epitopes for peptide vaccines.AVAILABILITY

AND IMPLEMENTATION: EpIC is available as a web server at

http://saphire.usask.ca/saphire/epic.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv136

PMCID: PMC4495294

PMID: 25750418 [Indexed for MEDLINE]

470. Bioinformatics. 2015 Jul 15;31(14):2276-83. doi: 10.1093/bioinformatics/btv133.

Epub 2015 Mar 3.

14-3-3-Pred: improved methods to predict 14-3-3-binding phosphopeptides.

Madeira F(1), Tinti M(2), Murugesan G(1), Berrett E(1), Stafford M(3), Toth R(4),

Cole C(1), MacKintosh C(2), Barton GJ(5).

Author information:

(1)Division of Computational Biology. (2)Division of Cell and Developmental

Biology. (3)MRC Protein Phosphorylation Unit. (4)Division of Signal Transduction

Therapy. (5)Division of Computational Biology, Division of Biological Chemistry

and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1

5EH, Scotland, UK.

MOTIVATION: The 14-3-3 family of phosphoprotein-binding proteins regulates many

cellular processes by docking onto pairs of phosphorylated Ser and Thr residues

in a constellation of intracellular targets. Therefore, there is a pressing need

to develop new prediction methods that use an updated set of 14-3-3-binding

motifs for the identification of new 14-3-3 targets and to prioritize the

downstream analysis of >2000 potential interactors identified in high-throughput

experiments.

RESULTS: Here, a comprehensive set of 14-3-3-binding targets from the literature

was used to develop 14-3-3-binding phosphosite predictors. Position-specific

scoring matrix, support vector machines (SVM) and artificial neural network (ANN)

classification methods were trained to discriminate experimentally determined

14-3-3-binding motifs from non-binding phosphopeptides. ANN, position-specific

scoring matrix and SVM methods showed best performance for a motif window

spanning from -6 to +4 around the binding phosphosite, achieving Matthews

correlation coefficient of up to 0.60. Blind prediction showed that all three

methods outperform two popular 14-3-3-binding site predictors, Scansite and ELM.

The new methods were used for prediction of 14-3-3-binding phosphosites in the

human proteome. Experimental analysis of high-scoring predictions in the FAM122A

and FAM122B proteins confirms the predictions and suggests the new

14-3-3-predictors will be generally useful.

AVAILABILITY AND IMPLEMENTATION: A standalone prediction web server is available

at http://www.compbio.dundee.ac.uk/1433pred. Human candidate 14-3-3-binding

phosphosites were integrated in ANIA: ANnotation and Integrated Analysis of the

14-3-3 interactome database.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv133

PMCID: PMC4495292

PMID: 25735772 [Indexed for MEDLINE]

471. PLoS One. 2015 Jul 15;10(7):e0133260. doi: 10.1371/journal.pone.0133260.

eCollection 2015.

SNBRFinder: A Sequence-Based Hybrid Algorithm for Enhanced Prediction of Nucleic

Acid-Binding Residues.

Yang X(1), Wang J(1), Sun J(1), Liu R(1).

Author information:

(1)Agricultural Bioinformatics Key Laboratory of Hubei Province, College of

Informatics, Huazhong Agricultural University, Wuhan, Hubei, People's Republic of

China.

Protein-nucleic acid interactions are central to various fundamental biological

processes. Automated methods capable of reliably identifying DNA- and RNA-binding

residues in protein sequence are assuming ever-increasing importance. The

majority of current algorithms rely on feature-based prediction, but their

accuracy remains to be further improved. Here we propose a sequence-based hybrid

algorithm SNBRFinder (Sequence-based Nucleic acid-Binding Residue Finder) by

merging a feature predictor SNBRFinderF and a template predictor SNBRFinderT.

SNBRFinderF was established using the support vector machine whose inputs include

sequence profile and other complementary sequence descriptors, while SNBRFinderT

was implemented with the sequence alignment algorithm based on profile hidden

Markov models to capture the weakly homologous template of query sequence.

Experimental results show that SNBRFinderF was clearly superior to the commonly

used sequence profile-based predictor and SNBRFinderT can achieve comparable

performance to the structure-based template methods. Leveraging the complementary

relationship between these two predictors, SNBRFinder reasonably improved the

performance of both DNA- and RNA-binding residue predictions. More importantly,

the sequence-based hybrid prediction reached competitive performance relative to

our previous structure-based counterpart. Our extensive and stringent comparisons

show that SNBRFinder has obvious advantages over the existing sequence-based

prediction algorithms. The value of our algorithm is highlighted by establishing

an easy-to-use web server that is freely accessible at

http://ibi.hzau.edu.cn/SNBRFinder.

DOI: 10.1371/journal.pone.0133260

PMCID: PMC4503397

PMID: 26176857 [Indexed for MEDLINE]

472. J Chem Theory Comput. 2015 Jul 14;11(7):2938-44. doi: 10.1021/acs.jctc.5b00190.

Epub 2015 Jun 2.

Properties of Organic Liquids when Simulated with Long-Range Lennard-Jones

Interactions.

Fischer NM(1), van Maaren PJ(1), Ditz JC(1), Yildirim A(2), van der Spoel D(1).

Author information:

(1)Uppsala Centre for Computational Chemistry, Science for Life Laboratory,

Department of Cell and Molecular Biology, Uppsala University , Husargatan 3, Box

596, SE-75124 Uppsala, Sweden. (2)Department of Physics, Faculty of Science and

Art, Siirt University , 56100 Siirt, Turkey.

In order to increase the accuracy of classical computer simulations, existing

methodologies may need to be adapted. Hitherto, most force fields employ a

truncated potential function to model van der Waals interactions, sometimes

augmented with an analytical correction. Although such corrections are accurate

for homogeneous systems with a long cutoff, they should not be used in inherently

inhomogeneous systems such as biomolecular and interface systems. For such cases,

a variant of the particle mesh Ewald algorithm (Lennard-Jones PME) was already

proposed 20 years ago (Essmann et al. J. Chem. Phys. 1995, 103, 8577-8593), but

it was implemented only recently (Wennberg et al. J. Chem. Theory Comput. 2013,

9, 3527-3537) in a major simulation code (GROMACS). The availability of this

method allows surface tensions of liquids as well as bulk properties to be

established, such as density and enthalpy of vaporization, without approximations

due to truncation. Here, we report on simulations of ≈150 liquids (taken from a

force field benchmark: Caleman et al. J. Chem. Theory Comput. 2012, 8, 61-74)

using three different force fields and compare simulations with and without

explicit long-range van der Waals interactions. We find that the density and

enthalpy of vaporization increase for most liquids using the generalized Amber

force field (GAFF, Wang et al. J. Comput. Chem. 2004, 25, 1157-1174) and the

Charmm generalized force field (CGenFF, Vanommeslaeghe et al. J. Comput. Chem.

2010, 31, 671-690) but less so for OPLS/AA (Jorgensen and Tirado-Rives, Proc.

Natl. Acad. Sci. U.S.A. 2005, 102, 6665-6670), which was parametrized with an

analytical correction to the van der Waals potential. The surface tension

increases by ≈10(-2) N/m for all force fields. These results suggest that van der

Waals attractions in force fields are too strong, in particular for the GAFF and

CGenFF. In addition to the simulation results, we introduce a new version of a

web server, http://virtualchemistry.org, aimed at facilitating sharing and reuse

of input files for molecular simulations.

DOI: 10.1021/acs.jctc.5b00190

PMID: 26575731 [Indexed for MEDLINE]

473. PLoS One. 2015 Jul 6;10(7):e0132305. doi: 10.1371/journal.pone.0132305.

eCollection 2015.

Basic Emotions in the Nencki Affective Word List (NAWL BE): New Method of

Classifying Emotional Stimuli.

Wierzba M(1), Riegel M(1), Wypych M(1), Jednoróg K(2), Turnau P(1), Grabowska

A(3), Marchewka A(1).

Author information:

(1)Laboratory of Brain Imaging, Neurobiology Centre, Nencki Institute of

Experimental Biology, Warsaw, Poland. (2)Laboratory of Psychophysiology,

Department of Neurophysiology, Nencki Institute of Experimental Biology, Warsaw,

Poland. (3)Laboratory of Psychophysiology, Department of Neurophysiology, Nencki

Institute of Experimental Biology, Warsaw, Poland; University of Social Sciences

and Humanities, Warsaw, Poland.

The Nencki Affective Word List (NAWL) has recently been introduced as a

standardized database of Polish words suitable for studying various aspects of

language and emotions. Though the NAWL was originally based on the most commonly

used dimensional approach, it is not the only way of studying emotions. Another

framework is based on discrete emotional categories. Since the two perspectives

are recognized as complementary, the aim of the present study was to supplement

the NAWL database by the addition of categories corresponding to basic emotions.

Thus, 2902 Polish words from the NAWL were presented to 265 subjects, who were

instructed to rate them according to the intensity of each of the five basic

emotions: happiness, anger, sadness, fear and disgust. The general

characteristics of the present word database, as well as the relationships

between the studied variables are shown to be consistent with typical patterns

found in previous studies using similar databases for different languages. Here

we present the Basic Emotions in the Nencki Affective Word List (NAWL BE) as a

database of verbal material suitable for highly controlled experimental research.

To make the NAWL more convenient to use, we introduce a comprehensive method of

classifying stimuli to basic emotion categories. We discuss the advantages of our

method in comparison to other methods of classification. Additionally, we provide

an interactive online tool (http://exp.lobi.nencki.gov.pl/nawl-analysis) to help

researchers browse and interactively generate classes of stimuli to meet their

specific requirements.

DOI: 10.1371/journal.pone.0132305

PMCID: PMC4492597

PMID: 26148193 [Indexed for MEDLINE]

474. Bioinformatics. 2015 Jul 1;31(13):2174-81. doi: 10.1093/bioinformatics/btv123.

Epub 2015 Feb 25.

Automated benchmarking of peptide-MHC class I binding predictions.

Trolle T(1), Metushi IG(1), Greenbaum JA(1), Kim Y(1), Sidney J(1), Lund O(1),

Sette A(1), Peters B(1), Nielsen M(2).

Author information:

(1)Center for Biological Sequence Analysis, Department of Systems Biology, The

Technical University of Denmark, Lyngby, Denmark, Division of Vaccine Discovery,

La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA and

Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín,

San Martín, Buenos Aires, Argentina. (2)Center for Biological Sequence Analysis,

Department of Systems Biology, The Technical University of Denmark, Lyngby,

Denmark, Division of Vaccine Discovery, La Jolla Institute for Allergy and

Immunology, La Jolla, CA 92037, USA and Instituto de Investigaciones

Biotecnológicas, Universidad Nacional de San Martín, San Martín, Buenos Aires,

Argentina Center for Biological Sequence Analysis, Department of Systems Biology,

The Technical University of Denmark, Lyngby, Denmark, Division of Vaccine

Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA

and Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San

Martín, San Martín, Buenos Aires, Argentina.

MOTIVATION: Numerous in silico methods predicting peptide binding to major

histocompatibility complex (MHC) class I molecules have been developed over the

last decades. However, the multitude of available prediction tools makes it

non-trivial for the end-user to select which tool to use for a given task. To

provide a solid basis on which to compare different prediction tools, we here

describe a framework for the automated benchmarking of peptide-MHC class I

binding prediction tools. The framework runs weekly benchmarks on data that are

newly entered into the Immune Epitope Database (IEDB), giving the public access

to frequent, up-to-date performance evaluations of all participating tools. To

overcome potential selection bias in the data included in the IEDB, a strategy

was implemented that suggests a set of peptides for which different prediction

methods give divergent predictions as to their binding capability. Upon

experimental binding validation, these peptides entered the benchmark study.

RESULTS: The benchmark has run for 15 weeks and includes evaluation of 44

datasets covering 17 MHC alleles and more than 4000 peptide-MHC binding

measurements. Inspection of the results allows the end-user to make educated

selections between participating tools. Of the four participating servers,

NetMHCpan performed the best, followed by ANN, SMM and finally ARB.

AVAILABILITY AND IMPLEMENTATION: Up-to-date performance evaluations of each

server can be found online at http://tools.iedb.org/auto\_bench/mhci/weekly. All

prediction tool developers are invited to participate in the benchmark. Sign-up

instructions are available at http://tools.iedb.org/auto\_bench/mhci/join.

CONTACT: mniel@cbs.dtu.dk or bpeters@liai.org

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv123

PMCID: PMC4481849

PMID: 25717196 [Indexed for MEDLINE]

475. Bioinformatics. 2015 Jul 1;31(13):2098-105. doi: 10.1093/bioinformatics/btv092.

Epub 2015 Feb 19.

AIDA: ab initio domain assembly for automated multi-domain protein structure

prediction and domain-domain interaction prediction.

Xu D(1), Jaroszewski L(2), Li Z(1), Godzik A(3).

Author information:

(1)Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research

Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA, Center for

Research in Biological Systems, University of California, San Diego, 9500 Gilman

Dr. La Jolla, CA 92093-0446, USA and Center of Excellence in Genomic Medicine

Research (CEGMR), King Fahad Medical Research Center, King Abdulaziz University,

Jeddah, Kingdom of Saudi Arabia. (2)Bioinformatics and Systems Biology Program,

Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La

Jolla, CA 92037, USA, Center for Research in Biological Systems, University of

California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093-0446, USA and Center of

Excellence in Genomic Medicine Research (CEGMR), King Fahad Medical Research

Center, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia Bioinformatics

and Systems Biology Program, Sanford-Burnham Medical Research Institute, 10901

North Torrey Pines Road, La Jolla, CA 92037, USA, Center for Research in

Biological Systems, University of California, San Diego, 9500 Gilman Dr. La

Jolla, CA 92093-0446, USA and Center of Excellence in Genomic Medicine Research

(CEGMR), King Fahad Medical Research Center, King Abdulaziz University, Jeddah,

Kingdom of Saudi Arabia. (3)Bioinformatics and Systems Biology Program,

Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La

Jolla, CA 92037, USA, Center for Research in Biological Systems, University of

California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093-0446, USA and Center of

Excellence in Genomic Medicine Research (CEGMR), King Fahad Medical Research

Center, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia Bioinformatics

and Systems Biology Program, Sanford-Burnham Medical Research Institute, 10901

North Torrey Pines Road, La Jolla, CA 92037, USA, Center for Research in

Biological Systems, University of California, San Diego, 9500 Gilman Dr. La

Jolla, CA 92093-0446, USA and Center of Excellence in Genomic Medicine Research

(CEGMR), King Fahad Medical Research Center, King Abdulaziz University, Jeddah,

Kingdom of Saudi Arabia Bioinformatics and Systems Biology Program,

Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La

Jolla, CA 92037, USA, Center for Research in Biological Systems, University of

California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093-0446, USA and Center of

Excellence in Genomic Medicine Research (CEGMR), King Fahad Medical Research

Center, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia.

MOTIVATION: Most proteins consist of multiple domains, independent structural and

evolutionary units that are often reshuffled in genomic rearrangements to form

new protein architectures. Template-based modeling methods can often detect

homologous templates for individual domains, but templates that could be used to

model the entire query protein are often not available.

RESULTS: We have developed a fast docking algorithm ab initio domain assembly

(AIDA) for assembling multi-domain protein structures, guided by the ab initio

folding potential. This approach can be extended to discontinuous domains (i.e.

domains with 'inserted' domains). When tested on experimentally solved structures

of multi-domain proteins, the relative domain positions were accurately found

among top 5000 models in 86% of cases. AIDA server can use domain assignments

provided by the user or predict them from the provided sequence. The latter

approach is particularly useful for automated protein structure prediction

servers. The blind test consisting of 95 CASP10 targets shows that domain

boundaries could be successfully determined for 97% of targets.

AVAILABILITY AND IMPLEMENTATION: The AIDA package as well as the benchmark sets

used here are available for download at http://ffas.burnham.org/AIDA/.

CONTACT: adam@sanfordburnham.org

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv092

PMCID: PMC4481839

PMID: 25701568 [Indexed for MEDLINE]

476. Bioinformatics. 2015 Jul 1;31(13):2091-7. doi: 10.1093/bioinformatics/btv091.

Epub 2015 Feb 16.

Folding RaCe: a robust method for predicting changes in protein folding rates

upon point mutations.

Chaudhary P(1), Naganathan AN(1), Gromiha MM(1).

Author information:

(1)Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences,

Indian Institute of Technology Madras, Chennai 600 036, India.

MOTIVATION: Protein engineering methods are commonly employed to decipher the

folding mechanism of proteins and enzymes. However, such experiments are

exceedingly time and resource intensive. It would therefore be advantageous to

develop a simple computational tool to predict changes in folding rates upon

mutations. Such a method should be able to rapidly provide the sequence position

and chemical nature to modulate through mutation, to effect a particular change

in rate. This can be of importance in protein folding, function or mechanistic

studies.

RESULTS: We have developed a robust knowledge-based methodology to predict the

changes in folding rates upon mutations formulated from amino and acid properties

using multiple linear regression approach. We benchmarked this method against an

experimental database of 790 point mutations from 26 two-state proteins. Mutants

were first classified according to secondary structure, accessible surface area

and position along the primary sequence. Three prime amino acid features

eliciting the best relationship with folding rates change were then shortlisted

for each class along with an optimized window length. We obtained a

self-consistent mean absolute error of 0.36 s(-1) and a mean Pearson correlation

coefficient (PCC) of 0.81. Jack-knife test resulted in a MAE of 0.42 s(-1) and a

PCC of 0.73. Moreover, our method highlights the importance of outlier(s)

detection and studying their implications in the folding mechanism.

AVAILABILITY AND IMPLEMENTATION: A web server 'Folding RaCe' has been developed

and is available at

http://www.iitm.ac.in/bioinfo/proteinfolding/foldingrace.html.

CONTACT: gromiha@iitm.ac.in

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv091

PMID: 25686635 [Indexed for MEDLINE]

477. Genome Biol Evol. 2015 Jul 1;7(6):1827-41. doi: 10.1093/gbe/evv106.

Reptilian Transcriptomes v2.0: An Extensive Resource for Sauropsida Genomics and

Transcriptomics.

Tzika AC(1), Ullate-Agote A(2), Grbic D(2), Milinkovitch MC(1).

Author information:

(1)Laboratory of Artificial & Natural Evolution (LANE), Department of Genetics &

Evolution, University of Geneva, Switzerland SIB Swiss Institute of

Bioinformatics, Switzerland Institute of Genetics and Genomics of Geneva (iGE3),

University of Geneva, Switzerland athanasia.tzika@unige.ch

michel.milinkovitch@unige.ch. (2)Laboratory of Artificial & Natural Evolution

(LANE), Department of Genetics & Evolution, University of Geneva, Switzerland SIB

Swiss Institute of Bioinformatics, Switzerland Institute of Genetics and Genomics

of Geneva (iGE3), University of Geneva, Switzerland.

Despite the availability of deep-sequencing techniques, genomic and

transcriptomic data remain unevenly distributed across phylogenetic groups. For

example, reptiles are poorly represented in sequence databases, hindering

functional evolutionary and developmental studies in these lineages substantially

more diverse than mammals. In addition, different studies use different assembly

and annotation protocols, inhibiting meaningful comparisons. Here, we present the

"Reptilian Transcriptomes Database 2.0," which provides extensive annotation of

transcriptomes and genomes from species covering the major reptilian lineages. To

this end, we sequenced normalized complementary DNA libraries of multiple adult

tissues and various embryonic stages of the leopard gecko and the corn snake and

gathered published reptilian sequence data sets from representatives of the four

extant orders of reptiles: Squamata (snakes and lizards), the tuatara,

crocodiles, and turtles. The LANE runner 2.0 software was implemented to annotate

all assemblies within a single integrated pipeline. We show that this approach

increases the annotation completeness of the assembled transcriptomes/genomes. We

then built large concatenated protein alignments of single-copy genes and

inferred phylogenetic trees that support the positions of turtles and the tuatara

as sister groups of Archosauria and Squamata, respectively. The Reptilian

Transcriptomes Database 2.0 resource will be updated to include selected new data

sets as they become available, thus making it a reference for differential

expression studies, comparative genomics and transcriptomics, linkage mapping,

molecular ecology, and phylogenomic analyses involving reptiles. The database is

available at www.reptilian-transcriptomes.org and can be enquired using a

wwwblast server installed at the University of Geneva.

© The Author(s) 2015. Published by Oxford University Press on behalf of the

Society for Molecular Biology and Evolution.

DOI: 10.1093/gbe/evv106

PMCID: PMC4494049

PMID: 26133641 [Indexed for MEDLINE]

478. J Biomol NMR. 2015 Jul;62(3):387-401. doi: 10.1007/s10858-015-9957-0. Epub 2015

Jun 16.

Accessible surface area from NMR chemical shifts.

Hafsa NE(1), Arndt D, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, Canada.

Accessible surface area (ASA) is the surface area of an atom, amino acid or

biomolecule that is exposed to solvent. The calculation of a molecule's ASA

requires three-dimensional coordinate data and the use of a "rolling ball"

algorithm to both define and calculate the ASA. For polymers such as proteins,

the ASA for individual amino acids is closely related to the hydrophobicity of

the amino acid as well as its local secondary and tertiary structure. For

proteins, ASA is a structural descriptor that can often be as informative as

secondary structure. Consequently there has been considerable effort over the

past two decades to try to predict ASA from protein sequence data and to use ASA

information (derived from chemical modification studies) as a structure

constraint. Recently it has become evident that protein chemical shifts are also

sensitive to ASA. Given the potential utility of ASA estimates as structural

constraints for NMR we decided to explore this relationship further. Using

machine learning techniques (specifically a boosted tree regression model) we

developed an algorithm called "ShiftASA" that combines chemical-shift and

sequence derived features to accurately estimate per-residue fractional ASA

values of water-soluble proteins. This method showed a correlation coefficient

between predicted and experimental values of 0.79 when evaluated on a set of 65

independent test proteins, which was an 8.2 % improvement over the next best

performing (sequence-only) method. On a separate test set of 92 proteins,

ShiftASA reported a mean correlation coefficient of 0.82, which was 12.3 % better

than the next best performing method. ShiftASA is available as a web server (

http://shiftasa.wishartlab.com ) for submitting input queries for fractional ASA

calculation.

DOI: 10.1007/s10858-015-9957-0

PMID: 26078090 [Indexed for MEDLINE]

479. Nucleic Acids Res. 2015 Jul 1;43(W1):W502-6. doi: 10.1093/nar/gkv557. Epub 2015

Jun 11.

RNAssess--a web server for quality assessment of RNA 3D structures.

Lukasiak P(1), Antczak M(2), Ratajczak T(2), Szachniuk M(3), Popenda M(4),

Adamiak RW(3), Blazewicz J(3).

Author information:

(1)Institute of Computing Science, Poznan University of Technology, Piotrowo 2,

60-965 Poznan, Poland Institute of Bioorganic Chemistry, Polish Academy of

Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland

piotr.lukasiak@cs.put.poznan.pl. (2)Institute of Computing Science, Poznan

University of Technology, Piotrowo 2, 60-965 Poznan, Poland. (3)Institute of

Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan,

Poland Institute of Bioorganic Chemistry, Polish Academy of Sciences,

Noskowskiego 12/14, 61-704 Poznan, Poland. (4)Institute of Bioorganic Chemistry,

Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland.

Nowadays, various methodologies can be applied to model RNA 3D structure. Thus,

the plausible quality assessment of 3D models has a fundamental impact on the

progress of structural bioinformatics. Here, we present RNAssess server, a novel

tool dedicated to visual evaluation of RNA 3D models in the context of the known

reference structure for a wide range of accuracy levels (from atomic to the whole

molecule perspective). The proposed server is based on the concept of local

neighborhood, defined as a set of atoms observed within a sphere localized around

a central atom of a particular residue. A distinctive feature of our server is

the ability to perform simultaneous visual analysis of the model-reference

structure coherence. RNAssess supports the quality assessment through delivering

both static and interactive visualizations that allows an easy identification of

native-like models and/or chosen structural regions of the analyzed molecule. A

combination of results provided by RNAssess allows us to rank analyzed models.

RNAssess offers new route to a fast and efficient 3D model evaluation suitable

for the RNA-Puzzles challenge. The proposed automated tool is implemented as a

free and open to all users web server with an user-friendly interface and can be

accessed at: http://rnassess.cs.put.poznan.pl/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv557

PMCID: PMC4489242

PMID: 26068469 [Indexed for MEDLINE]

480. Nucleic Acids Res. 2015 Jul 1;43(W1):W15-23. doi: 10.1093/nar/gkv543. Epub 2015

Jun 5.

R3D-2-MSA: the RNA 3D structure-to-multiple sequence alignment server.

Cannone JJ(1), Sweeney BA(2), Petrov AI(3), Gutell RR(1), Zirbel CL(4), Leontis

N(5).

Author information:

(1)Institute for Cellular and Molecular Biology, and the Center for Computational

Biology and Bioinformatics, The University of Texas at Austin, Austin, TX 78712,

USA. (2)Department of Biological Sciences, Bowling Green State University,

Bowling Green, OH 43403, USA. (3)European Molecular Biology Laboratory, European

Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10

1SD, UK. (4)Department of Mathematics and Statistics, Bowling Green State

University, Bowling Green, OH 43403, USA. (5)Department of Chemistry, Bowling

Green State University, Bowling Green, OH 43403, USA leontis@bgsu.edu.

The RNA 3D Structure-to-Multiple Sequence Alignment Server (R3D-2-MSA) is a new

web service that seamlessly links RNA three-dimensional (3D) structures to

high-quality RNA multiple sequence alignments (MSAs) from diverse biological

sources. In this first release, R3D-2-MSA provides manual and programmatic access

to curated, representative ribosomal RNA sequence alignments from bacterial,

archaeal, eukaryal and organellar ribosomes, using nucleotide numbers from

representative atomic-resolution 3D structures. A web-based front end is

available for manual entry and an Application Program Interface for programmatic

access. Users can specify up to five ranges of nucleotides and 50 nucleotide

positions per range. The R3D-2-MSA server maps these ranges to the appropriate

columns of the corresponding MSA and returns the contents of the columns, either

for display in a web browser or in JSON format for subsequent programmatic use.

The browser output page provides a 3D interactive display of the query, a full

list of sequence variants with taxonomic information and a statistical summary of

distinct sequence variants found. The output can be filtered and sorted in the

browser. Previous user queries can be viewed at any time by resubmitting the

output URL, which encodes the search and re-generates the results. The service is

freely available with no login requirement at http://rna.bgsu.edu/r3d-2-msa.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv543

PMCID: PMC4489222

PMID: 26048960 [Indexed for MEDLINE]

481. Nucleic Acids Res. 2015 Jul 1;43(W1):W213-9. doi: 10.1093/nar/gkv404. Epub 2015

Jun 4.

CATNAP: a tool to compile, analyze and tally neutralizing antibody panels.

Yoon H(1), Macke J(1), West AP Jr(2), Foley B(1), Bjorkman PJ(2), Korber B(1),

Yusim K(3).

Author information:

(1)Los Alamos National Laboratory, Los Alamos, NM, USA. (2)California Institute

of Technology, Pasadena, CA, USA. (3)Los Alamos National Laboratory, Los Alamos,

NM, USA kyusim@lanl.gov.

CATNAP (Compile, Analyze and Tally NAb Panels) is a new web server at Los Alamos

HIV Database, created to respond to the newest advances in HIV neutralizing

antibody research. It is a comprehensive platform focusing on neutralizing

antibody potencies in conjunction with viral sequences. CATNAP integrates

neutralization and sequence data from published studies, and allows users to

analyze that data for each HIV Envelope protein sequence position and each

antibody. The tool has multiple data retrieval and analysis options. As input,

the user can pick specific antibodies and viruses, choose a panel from a

published study, or supply their own data. The output superimposes neutralization

panel data, virus epidemiological data, and viral protein sequence alignments on

one page, and provides further information and analyses. The user can highlight

alignment positions, or select antibody contact residues and view

position-specific information from the HIV databases. The tool calculates tallies

of amino acids and N-linked glycosylation motifs, counts of antibody-sensitive

and -resistant viruses in conjunction with each amino acid or N-glycosylation

motif, and performs Fisher's exact test to detect potential positive or negative

amino acid associations for the selected antibody. Website name: CATNAP (Compile,

Analyze and Tally NAb Panels). Website address: http://hiv.lanl.gov/catnap.

Published by Oxford University Press on behalf of Nucleic Acids Research

2015.This work is written by a US Government employee and is in the public domain

in the US.

DOI: 10.1093/nar/gkv404

PMCID: PMC4489231

PMID: 26044712 [Indexed for MEDLINE]

482. Nucleic Acids Res. 2015 Jul 1;43(W1):W134-40. doi: 10.1093/nar/gkv523. Epub 2015

May 27.

INGA: protein function prediction combining interaction networks, domain

assignments and sequence similarity.

Piovesan D(1), Giollo M(2), Leonardi E(3), Ferrari C(4), Tosatto SC(5).

Author information:

(1)Department of Biomedical Sciences, University of Padua, Padua 35121, Italy.

(2)Department of Biomedical Sciences, University of Padua, Padua 35121, Italy

Department of Information Engineering, University of Padua, Padua 35121, Italy.

(3)Department of Women's and Children's Health, University of Padua, Padua 35128,

Italy. (4)Department of Information Engineering, University of Padua, Padua

35121, Italy. (5)Department of Biomedical Sciences, University of Padua, Padua

35121, Italy CNR Institute of Neuroscience, Padua 35121, Italy

silvio.tosatto@unipd.it.

Identifying protein functions can be useful for numerous applications in biology.

The prediction of gene ontology (GO) functional terms from sequence remains

however a challenging task, as shown by the recent CAFA experiments. Here we

present INGA, a web server developed to predict protein function from a

combination of three orthogonal approaches. Sequence similarity and domain

architecture searches are combined with protein-protein interaction network data

to derive consensus predictions for GO terms using functional enrichment. The

INGA server can be queried both programmatically through RESTful services and

through a web interface designed for usability. The latter provides output

supporting the GO term predictions with the annotating sequences. INGA is

validated on the CAFA-1 data set and was recently shown to perform consistently

well in the CAFA-2 blind test. The INGA web server is available from URL:

http://protein.bio.unipd.it/inga.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv523

PMCID: PMC4489281

PMID: 26019177 [Indexed for MEDLINE]

483. Nucleic Acids Res. 2015 Jul 1;43(W1):W513-21. doi: 10.1093/nar/gkv460. Epub 2015

May 27.

RNAiFold 2.0: a web server and software to design custom and Rfam-based RNA

molecules.

Garcia-Martin JA(1), Dotu I(2), Clote P(3).

Author information:

(1)Biology Department, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA

02467, USA. (2)Research Programme on Biomedical Informatics (GRIB), Department of

Experimental and Health Sciences, Universitat Pompeu Fabra, IMIM (Hospital del

Mar Medical Research Institute); C/Dr. Aiguader 88, Barcelona E-08003, Spain.

(3)Biology Department, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA

02467, USA clote@bc.edu.

Several algorithms for RNA inverse folding have been used to design synthetic

riboswitches, ribozymes and thermoswitches, whose activity has been

experimentally validated. The RNAiFold software is unique among approaches for

inverse folding in that (exhaustive) constraint programming is used instead of

heuristic methods. For that reason, RNAiFold can generate all sequences that fold

into the target structure or determine that there is no solution. RNAiFold 2.0 is

a complete overhaul of RNAiFold 1.0, rewritten from the now defunct COMET

language to C++. The new code properly extends the capabilities of its

predecessor by providing a user-friendly pipeline to design synthetic constructs

having the functionality of given Rfam families. In addition, the new software

supports amino acid constraints, even for proteins translated in different

reading frames from overlapping coding sequences; moreover, structure

compatibility/incompatibility constraints have been expanded. With these

features, RNAiFold 2.0 allows the user to design single RNA molecules as well as

hybridization complexes of two RNA molecules.AVAILABILITY: the web server, source

code and linux binaries are publicly accessible at

http://bioinformatics.bc.edu/clotelab/RNAiFold2.0.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv460

PMCID: PMC4489274

PMID: 26019176 [Indexed for MEDLINE]

484. Nucleic Acids Res. 2015 Jul 1;43(W1):W289-94. doi: 10.1093/nar/gkv556. Epub 2015

May 26.

DeAnnCNV: a tool for online detection and annotation of copy number variations

from whole-exome sequencing data.

Zhang Y(1), Yu Z(2), Ban R(2), Zhang H(1), Iqbal F(3), Zhao A(4), Li A(5), Shi

Q(6).

Author information:

(1)Molecular and Cell Genetics Laboratory, The CAS Key Laboratory of Innate

Immunity and Chronic Disease, Hefei National Laboratory for Physical Sciences at

Microscale and School of Life Sciences, University of Science and Technology of

China, Hefei 230027, China. (2)School of Information Science and Technology,

University of Science and Technology of China, Hefei 230027, China. (3)Molecular

and Cell Genetics Laboratory, The CAS Key Laboratory of Innate Immunity and

Chronic Disease, Hefei National Laboratory for Physical Sciences at Microscale

and School of Life Sciences, University of Science and Technology of China, Hefei

230027, China Institute of Pure and Applied Biology, Bahauddin Zakariya

University Multan, 60800, Pakistan. (4)Hefei Institute of Physical Science, China

Academy of Science, Hefei 230027, China. (5)School of Information Science and

Technology, University of Science and Technology of China, Hefei 230027, China

Research Centers for Biomedical Engineering, University of Science and Technology

of China, Hefei 230027, China aoli@ustc.edu.cn. (6)Molecular and Cell Genetics

Laboratory, The CAS Key Laboratory of Innate Immunity and Chronic Disease, Hefei

National Laboratory for Physical Sciences at Microscale and School of Life

Sciences, University of Science and Technology of China, Hefei 230027, China

qshi@ustc.edu.cn.

With the decrease in costs, whole-exome sequencing (WES) has become a very

popular and powerful tool for the identification of genetic variants underlying

human diseases. However, integrated tools to precisely detect and systematically

annotate copy number variations (CNVs) from WES data are still in great demand.

Here, we present an online tool, DeAnnCNV (Detection and Annotation of Copy

Number Variations from WES data), to meet the current demands of WES users. Upon

submitting the file generated from WES data by an in-house tool that can be

downloaded from our server, DeAnnCNV can detect CNVs in each sample and extract

the shared CNVs among multiple samples. DeAnnCNV also provides additional useful

supporting information for the detected CNVs and associated genes to help users

to find the potential candidates for further experimental study. The web server

is implemented in PHP + Perl + MATLAB and is online available to all users for

free at http://mcg.ustc.edu.cn/db/cnv/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv556

PMCID: PMC4489280

PMID: 26013811 [Indexed for MEDLINE]

485. Nucleic Acids Res. 2015 Jul 1;43(W1):W349-55. doi: 10.1093/nar/gkv535. Epub 2015

May 24.

LYRA, a webserver for lymphocyte receptor structural modeling.

Klausen MS(1), Anderson MV(1), Jespersen MC(1), Nielsen M(2), Marcatili P(3).

Author information:

(1)Center for Biological Sequence Analysis, Technical University of Denmark, Kgs.

Lyngby, Denmark. (2)Center for Biological Sequence Analysis, Technical University

of Denmark, Kgs. Lyngby, Denmark Instituto de Investigaciones Biotecnológicas,

Universidad Nacional de San Martín, Buenos Aires, Argentina mniel@cbs.dtu.dk.

(3)Center for Biological Sequence Analysis, Technical University of Denmark, Kgs.

Lyngby, Denmark paolo.marcatili@gmail.com.

The accurate structural modeling of B- and T-cell receptors is fundamental to

gain a detailed insight in the mechanisms underlying immunity and in developing

new drugs and therapies. The LYRA (LYmphocyte Receptor Automated modeling) web

server (http://www.cbs.dtu.dk/services/LYRA/) implements a complete and automated

method for building of B- and T-cell receptor structural models starting from

their amino acid sequence alone. The webserver is freely available and easy to

use for non-specialists. Upon submission, LYRA automatically generates alignments

using ad hoc profiles, predicts the structural class of each hypervariable loop,

selects the best templates in an automatic fashion, and provides within minutes a

complete 3D model that can be downloaded or inspected online. Experienced users

can manually select or exclude template structures according to case specific

information. LYRA is based on the canonical structure method, that in the last 30

years has been successfully used to generate antibody models of high accuracy,

and in our benchmarks this approach proves to achieve similarly good results on

TCR modeling, with a benchmarked average RMSD accuracy of 1.29 and 1.48 Å for B-

and T-cell receptors, respectively. To the best of our knowledge, LYRA is the

first automated server for the prediction of TCR structure.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv535

PMCID: PMC4489227

PMID: 26007650 [Indexed for MEDLINE]

486. Nucleic Acids Res. 2015 Jul 1;43(W1):W522-6. doi: 10.1093/nar/gkv538. Epub 2015

May 20.

Primerize: automated primer assembly for transcribing non-coding RNA domains.

Tian S(1), Yesselman JD(1), Cordero P(2), Das R(3).

Author information:

(1)Departments of Biochemistry, Stanford University, Stanford CA 94305, USA.

(2)Program in Biomedical Informatics, Stanford University, Stanford CA 94305,

USA. (3)Departments of Biochemistry, Stanford University, Stanford CA 94305, USA

Program in Biomedical Informatics, Stanford University, Stanford CA 94305, USA

Department of Physics, Stanford University, Stanford CA 94305, USA

rhiju@stanford.edu.

Customized RNA synthesis is in demand for biological and biotechnological

research. While chemical synthesis and gel or chromatographic purification of RNA

is costly and difficult for sequences longer than tens of nucleotides, a pipeline

of primer assembly of DNA templates, in vitro transcription by T7 RNA polymerase

and kit-based purification provides a cost-effective and fast alternative for

preparing RNA molecules. Nevertheless, designing template primers that optimize

cost and avoid mispriming during polymerase chain reaction currently requires

expert inspection, downloading specialized software or both. Online servers are

currently not available or maintained for the task. We report here a server named

Primerize that makes available an efficient algorithm for primer design developed

and experimentally tested in our laboratory for RNA domains with lengths up to

300 nucleotides. Free access: http://primerize.stanford.edu.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv538

PMCID: PMC4489279

PMID: 25999345 [Indexed for MEDLINE]

487. Nucleic Acids Res. 2015 Jul 1;43(W1):W258-63. doi: 10.1093/nar/gkv515. Epub 2015

May 18.

MyProteinNet: build up-to-date protein interaction networks for organisms,

tissues and user-defined contexts.

Basha O(1), Flom D(2), Barshir R(1), Smoly I(2), Tirman S(1), Yeger-Lotem E(3).

Author information:

(1)Department of Clinical Biochemistry & Pharmacology, Faculty of Health

Sciences, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel.

(2)Department of Computer Science, Faculty of Natural Sciences, Ben-Gurion

University of the Negev, Beer-Sheva 84105, Israel. (3)Department of Clinical

Biochemistry & Pharmacology, Faculty of Health Sciences, Ben-Gurion University of

the Negev, Beer-Sheva 84105, Israel National Institute for Biotechnology in the

Negev, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

estiyl@bgu.ac.il.

The identification of the molecular pathways active in specific contexts, such as

disease states or drug responses, often requires an extensive view of the

potential interactions between a subset of proteins. This view is not easily

obtained: it requires the integration of context-specific protein list or

expression data with up-to-date data of protein interactions that are typically

spread across multiple databases. The MyProteinNet web server allows users to

easily create such context-sensitive protein interaction networks. Users can

automatically gather and consolidate data from up to 11 different databases to

create a generic protein interaction network (interactome). They can score the

interactions based on reliability and filter them by user-defined contexts

including molecular expression and protein annotation. The output of MyProteinNet

includes the generic and filtered interactome files, together with a summary of

their network attributes. MyProteinNet is particularly geared toward building

human tissue interactomes, by maintaining tissue expression profiles from

multiple resources. The ability of MyProteinNet to facilitate the construction of

up-to-date, context-specific interactomes and its applicability to 11 different

organisms and to tens of human tissues, make it a powerful tool in meaningful

analysis of protein networks. MyProteinNet is available at

http://netbio.bgu.ac.il/myproteinnet.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv515

PMCID: PMC4489290

PMID: 25990735 [Indexed for MEDLINE]

488. Nucleic Acids Res. 2015 Jul 1;43(W1):W480-6. doi: 10.1093/nar/gkv524. Epub 2015

May 18.

StarScan: a web server for scanning small RNA targets from degradome sequencing

data.

Liu S(1), Li JH(1), Wu J(1), Zhou KR(1), Zhou H(1), Yang JH(2), Qu LH(2).

Author information:

(1)Key Laboratory of Gene Engineering of the Ministry of Education, Sun Yat-sen

University, Guangzhou 510275, P. R. China. (2)Key Laboratory of Gene Engineering

of the Ministry of Education, Sun Yat-sen University, Guangzhou 510275, P. R.

China State Key Laboratory for Biocontrol, Sun Yat-sen University, Guangzhou

510275, P. R. China lssqlh@mail.sysu.edu.cn.

Endogenous small non-coding RNAs (sRNAs), including microRNAs, PIWI-interacting

RNAs and small interfering RNAs, play important gene regulatory roles in animals

and plants by pairing to the protein-coding and non-coding transcripts. However,

computationally assigning these various sRNAs to their regulatory target genes

remains technically challenging. Recently, a high-throughput degradome sequencing

method was applied to identify biologically relevant sRNA cleavage sites. In this

study, an integrated web-based tool, StarScan (sRNA target Scan), was developed

for scanning sRNA targets using degradome sequencing data from 20 species. Given

a sRNA sequence from plants or animals, our web server performs an ultrafast and

exhaustive search for potential sRNA-target interactions in annotated and

unannotated genomic regions. The interactions between small RNAs and target

transcripts were further evaluated using a novel tool, alignScore. A novel tool,

degradomeBinomTest, was developed to quantify the abundance of degradome

fragments located at the 9-11th nucleotide from the sRNA 5' end. This is the

first web server for discovering potential sRNA-mediated RNA cleavage events in

plants and animals, which affords mechanistic insights into the regulatory roles

of sRNAs. The StarScan web server is available at

http://mirlab.sysu.edu.cn/starscan/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv524

PMCID: PMC4489260

PMID: 25990732 [Indexed for MEDLINE]

489. Nucleic Acids Res. 2015 Jul 1;43(W1):W326-30. doi: 10.1093/nar/gkv542. Epub 2015

May 18.

MS2PIP prediction server: compute and visualize MS2 peak intensity predictions

for CID and HCD fragmentation.

Degroeve S(1), Maddelein D(2), Martens L(2).

Author information:

(1)Department of Medical Protein Research, VIB, A. Baertsoenkaai 3, B-9000 Ghent,

Belgium Department of Biochemistry, Faculty of Medicine and Health Sciences,

Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium

sven.degroeve@vib-ugent.be. (2)Department of Medical Protein Research, VIB, A.

Baertsoenkaai 3, B-9000 Ghent, Belgium Department of Biochemistry, Faculty of

Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent,

Belgium.

We present an MS(2) peak intensity prediction server that computes MS(2) charge

2+ and 3+ spectra from peptide sequences for the most common fragment ions. The

server integrates the Unimod public domain post-translational modification

database for modified peptides. The prediction model is an improvement of the

previously published MS(2)PIP model for Orbitrap-LTQ CID spectra. Predicted MS(2)

spectra can be downloaded as a spectrum file and can be visualized in the browser

for comparisons with observations. In addition, we added prediction models for

HCD fragmentation (Q-Exactive Orbitrap) and show that these models compute

accurate intensity predictions on par with CID performance. We also show that

training prediction models for CID and HCD separately improves the accuracy for

each fragmentation method. The MS(2)PIP prediction server is accessible from

http://iomics.ugent.be/ms2pip.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv542

PMCID: PMC4489309

PMID: 25990723 [Indexed for MEDLINE]

490. Nucleic Acids Res. 2015 Jul 1;43(W1):W370-7. doi: 10.1093/nar/gkv494. Epub 2015

May 15.

CSI 3.0: a web server for identifying secondary and super-secondary structure in

proteins using NMR chemical shifts.

Hafsa NE(1), Arndt D(1), Wishart DS(2).

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8,

Canada. (2)Department of Computing Science, University of Alberta, Edmonton, AB

T6G 2E8, Canada Department of Biological Sciences, University of Alberta,

Edmonton, AB T6G 2E8, Canada david.wishart@ualberta.ca.

The Chemical Shift Index or CSI 3.0 (http://csi3.wishartlab.com) is a web server

designed to accurately identify the location of secondary and super-secondary

structures in protein chains using only nuclear magnetic resonance (NMR) backbone

chemical shifts and their corresponding protein sequence data. Unlike earlier

versions of CSI, which only identified three types of secondary structure (helix,

β-strand and coil), CSI 3.0 now identifies total of 11 types of secondary and

super-secondary structures, including helices, β-strands, coil regions, five

common β-turns (type I, II, I', II' and VIII), β hairpins as well as interior and

edge β-strands. CSI 3.0 accepts experimental NMR chemical shift data in multiple

formats (NMR Star 2.1, NMR Star 3.1 and SHIFTY) and generates colorful CSI plots

(bar graphs) and secondary/super-secondary structure assignments. The output can

be readily used as constraints for structure determination and refinement or the

images may be used for presentations and publications. CSI 3.0 uses a pipeline of

several well-tested, previously published programs to identify the secondary and

super-secondary structures in protein chains. Comparisons with secondary and

super-secondary structure assignments made via standard coordinate analysis

programs such as DSSP, STRIDE and VADAR on high-resolution protein structures

solved by X-ray and NMR show >90% agreement between those made with CSI 3.0.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv494

PMCID: PMC4489240

PMID: 25979265 [Indexed for MEDLINE]

491. Nucleic Acids Res. 2015 Jul 1;43(W1):W141-7. doi: 10.1093/nar/gkv461. Epub 2015

May 15.

SIFTER search: a web server for accurate phylogeny-based protein function

prediction.

Sahraeian SM(1), Luo KR(2), Brenner SE(3).

Author information:

(1)Department of Plant and Microbial Biology, University of California, Berkeley,

CA 94720, USA. (2)Department of Molecular and Cell Biology, University of

California, Berkeley, CA 94720, USA. (3)Department of Plant and Microbial

Biology, University of California, Berkeley, CA 94720, USA Department of

Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA

Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley,

CA 94720, USA brenner@compbio.berkeley.edu.

We are awash in proteins discovered through high-throughput sequencing projects.

As only a minuscule fraction of these have been experimentally characterized,

computational methods are widely used for automated annotation. Here, we

introduce a user-friendly web interface for accurate protein function prediction

using the SIFTER algorithm. SIFTER is a state-of-the-art sequence-based gene

molecular function prediction algorithm that uses a statistical model of function

evolution to incorporate annotations throughout the phylogenetic tree. Due to the

resources needed by the SIFTER algorithm, running SIFTER locally is not trivial

for most users, especially for large-scale problems. The SIFTER web server thus

provides access to precomputed predictions on 16 863 537 proteins from 232 403

species. Users can explore SIFTER predictions with queries for proteins, species,

functions, and homologs of sequences not in the precomputed prediction set. The

SIFTER web server is accessible at http://sifter.berkeley.edu/ and the source

code can be downloaded.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv461

PMCID: PMC4489292

PMID: 25979264 [Indexed for MEDLINE]

492. Nucleic Acids Res. 2015 Jul 1;43(W1):W331-7. doi: 10.1093/nar/gkv490. Epub 2015

May 14.

PrionW: a server to identify proteins containing glutamine/asparagine rich

prion-like domains and their amyloid cores.

Zambrano R(1), Conchillo-Sole O(1), Iglesias V(1), Illa R(1), Rousseau F(2),

Schymkowitz J(2), Sabate R(3), Daura X(4), Ventura S(5).

Author information:

(1)Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i

Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain.

(2)VIB Switch Laboratory and Department for Cellular and Molecular Medicine, KU

Leuven, Leuven, Belgium. (3)Institut de Nanociència i Nanotecnologia (INUB) and

Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona,

Barcelona, Spain. (4)Institut de Biotecnologia i de Biomedicina and Departament

de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona,

Bellaterra, 08193, Spain Institució Catalana de Recerca i Estudis Avançats

(ICREA), Barcelona, Spain. (5)Institut de Biotecnologia i de Biomedicina and

Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de

Barcelona, Bellaterra, 08193, Spain salvador.ventura@uab.es.

Prions are a particular type of amyloids with the ability to self-perpetuate and

propagate in vivo. Prion-like conversion underlies important biological processes

but is also connected to human disease. Yeast prions are the best understood

transmissible amyloids. In these proteins, prion formation from an initially

soluble state involves a structural conversion, driven, in many cases, by

specific domains enriched in glutamine/asparagine (Q/N) residues. Importantly,

domains sharing this compositional bias are also present in the proteomes of

higher organisms, thus suggesting that prion-like conversion might be an

evolutionary conserved mechanism. We have recently shown that the identification

and evaluation of the potency of amyloid nucleating sequences in putative prion

domains allows discrimination of genuine prions. PrionW is a web application that

exploits this principle to scan sequences in order to identify proteins

containing Q/N enriched prion-like domains (PrLDs) in large datasets. When used

to scan the complete yeast proteome, PrionW identifies previously experimentally

validated prions with high accuracy. Users can analyze up to 10 000 sequences at

a time, PrLD-containing proteins are identified and their putative PrLDs and

amyloid nucleating cores visualized and scored. The output files can be

downloaded for further analysis. PrionW server can be accessed at

http://bioinf.uab.cat/prionw/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv490

PMCID: PMC4489250

PMID: 25977297 [Indexed for MEDLINE]

493. Nucleic Acids Res. 2015 Jul 1;43(W1):W425-30. doi: 10.1093/nar/gkv493. Epub 2015

May 14.

NPDock: a web server for protein-nucleic acid docking.

Tuszynska I(1), Magnus M(2), Jonak K(2), Dawson W(2), Bujnicki JM(3).

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, PL-02-109 Warsaw,

Poland Institute of Informatics, University of Warsaw, Banacha 2, PL-02-097

Warsaw, Poland. (2)Laboratory of Bioinformatics and Protein Engineering,

International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena

4, PL-02-109 Warsaw, Poland. (3)Laboratory of Bioinformatics and Protein

Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul.

Ks. Trojdena 4, PL-02-109 Warsaw, Poland Bioinformatics Laboratory, Institute of

Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz

University, ul. Umultowska 89, PL-61-614 Poznan, Poland iamb@genesilico.pl.

Protein-RNA and protein-DNA interactions play fundamental roles in many

biological processes. A detailed understanding of these interactions requires

knowledge about protein-nucleic acid complex structures. Because the experimental

determination of these complexes is time-consuming and perhaps futile in some

instances, we have focused on computational docking methods starting from the

separate structures. Docking methods are widely employed to study protein-protein

interactions; however, only a few methods have been made available to model

protein-nucleic acid complexes. Here, we describe NPDock (Nucleic acid-Protein

Docking); a novel web server for predicting complexes of protein-nucleic acid

structures which implements a computational workflow that includes docking,

scoring of poses, clustering of the best-scored models and refinement of the most

promising solutions. The NPDock server provides a user-friendly interface and 3D

visualization of the results. The smallest set of input data consists of a

protein structure and a DNA or RNA structure in PDB format. Advanced options are

available to control specific details of the docking process and obtain

intermediate results. The web server is available at http://genesilico.pl/NPDock.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv493

PMCID: PMC4489298

PMID: 25977296 [Indexed for MEDLINE]

494. Nucleic Acids Res. 2015 Jul 1;43(W1):W460-6. doi: 10.1093/nar/gkv403. Epub 2015

May 14.

DIANA-miRPath v3.0: deciphering microRNA function with experimental support.

Vlachos IS(1), Zagganas K(2), Paraskevopoulou MD(3), Georgakilas G(3), Karagkouni

D(3), Vergoulis T(4), Dalamagas T(5), Hatzigeorgiou AG(6).

Author information:

(1)DIANA-Lab, Department of Electrical & Computer Engineering, University of

Thessaly, 382 21 Volos, Greece Laboratory for Experimental Surgery and Surgical

Research 'N.S. Christeas', Medical School of Athens, University of Athens, 11527

Athens, Greece ivlachos@lessr.eu. (2)'Athena' Research and Innovation Center,

11524 Athens, Greece University of Peloponnisos, Department of Informatics and

Telecommunications, 22100 Tripoli, Greece. (3)DIANA-Lab, Department of Electrical

& Computer Engineering, University of Thessaly, 382 21 Volos, Greece. (4)'Athena'

Research and Innovation Center, 11524 Athens, Greece. (5)'Athena' Research and

Innovation Center, 11524 Athens, Greece ivlachos@lessr.eu. (6)DIANA-Lab,

Department of Electrical & Computer Engineering, University of Thessaly, 382 21

Volos, Greece 'Athena' Research and Innovation Center, 11524 Athens, Greece

ivlachos@lessr.eu.

The functional characterization of miRNAs is still an open challenge. Here, we

present DIANA-miRPath v3.0 (http://www.microrna.gr/miRPathv3) an online software

suite dedicated to the assessment of miRNA regulatory roles and the

identification of controlled pathways. The new miRPath web server renders

possible the functional annotation of one or more miRNAs using standard

(hypergeometric distributions), unbiased empirical distributions and/or

meta-analysis statistics. DIANA-miRPath v3.0 database and functionality have been

significantly extended to support all analyses for KEGG molecular pathways, as

well as multiple slices of Gene Ontology (GO) in seven species (Homo sapiens, Mus

musculus, Rattus norvegicus, Drosophila melanogaster, Caenorhabditis elegans,

Gallus gallus and Danio rerio). Importantly, more than 600 000 experimentally

supported miRNA targets from DIANA-TarBase v7.0 have been incorporated into the

new schema. Users of DIANA-miRPath v3.0 can harness this wealth of information

and substitute or combine the available in silico predicted targets from

DIANA-microT-CDS and/or TargetScan v6.2 with high quality experimentally

supported interactions. A unique feature of DIANA-miRPath v3.0 is its redesigned

Reverse Search module, which enables users to identify and visualize miRNAs

significantly controlling selected pathways or belonging to specific GO

categories based on in silico or experimental data. DIANA-miRPath v3.0 is freely

available to all users without any login requirement.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv403

PMCID: PMC4489228

PMID: 25977294 [Indexed for MEDLINE]

495. Nucleic Acids Res. 2015 Jul 1;43(W1):W493-7. doi: 10.1093/nar/gkv489. Epub 2015

May 14.

Web-Beagle: a web server for the alignment of RNA secondary structures.

Mattei E(1), Pietrosanto M(1), Ferrè F(2), Helmer-Citterich M(1).

Author information:

(1)Centre for Molecular Bioinformatics, Department of Biology, University of Rome

Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy. (2)Department

of Pharmacy and Biotechnology (FaBiT), University of Bologna Alma Mater, Via

Belmeloro 6, 40126 Bologna, Italy fabrizio.ferre@unibo.it.

Web-Beagle (http://beagle.bio.uniroma2.it) is a web server for the pairwise

global or local alignment of RNA secondary structures. The server exploits a new

encoding for RNA secondary structure and a substitution matrix of RNA structural

elements to perform RNA structural alignments. The web server allows the user to

compute up to 10 000 alignments in a single run, taking as input sets of RNA

sequences and structures or primary sequences alone. In the latter case, the

server computes the secondary structure prediction for the RNAs on-the-fly using

RNAfold (free energy minimization). The user can also compare a set of input RNAs

to one of five pre-compiled RNA datasets including lncRNAs and 3' UTRs. All types

of comparison produce in output the pairwise alignments along with structural

similarity and statistical significance measures for each resulting alignment. A

graphical color-coded representation of the alignments allows the user to easily

identify structural similarities between RNAs. Web-Beagle can be used for finding

structurally related regions in two or more RNAs, for the identification of

homologous regions or for functional annotation. Benchmark tests show that

Web-Beagle has lower computational complexity, running time and better

performances than other available methods.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv489

PMCID: PMC4489221

PMID: 25977293 [Indexed for MEDLINE]

496. Nucleic Acids Res. 2015 Jul 1;43(W1):W128-33. doi: 10.1093/nar/gkv486. Epub 2015

May 12.

IMP 2.0: a multi-species functional genomics portal for integration,

visualization and prediction of protein functions and networks.

Wong AK(1), Krishnan A(2), Yao V(3), Tadych A(2), Troyanskaya OG(4).

Author information:

(1)Department of Computer Science, Princeton University, Princeton, NJ 08540, USA

Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton,

NJ 08540, USA Simons Center for Data Analysis, Simons Foundation, NY 10010, USA.

(2)Lewis-Sigler Institute for Integrative Genomics, Princeton University,

Princeton, NJ 08540, USA. (3)Department of Computer Science, Princeton

University, Princeton, NJ 08540, USA Lewis-Sigler Institute for Integrative

Genomics, Princeton University, Princeton, NJ 08540, USA. (4)Department of

Computer Science, Princeton University, Princeton, NJ 08540, USA Lewis-Sigler

Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540,

USA Simons Center for Data Analysis, Simons Foundation, NY 10010, USA

ogt@cs.princeton.edu.

IMP (Integrative Multi-species Prediction), originally released in 2012, is an

interactive web server that enables molecular biologists to interpret

experimental results and to generate hypotheses in the context of a large

cross-organism compendium of functional predictions and networks. The system

provides biologists with a framework to analyze their candidate gene sets in the

context of functional networks, expanding or refining their sets using functional

relationships predicted from integrated high-throughput data. IMP 2.0 integrates

updated prior knowledge and data collections from the last three years in the

seven supported organisms (Homo sapiens, Mus musculus, Rattus norvegicus,

Drosophila melanogaster, Danio rerio, Caenorhabditis elegans, and Saccharomyces

cerevisiae) and extends function prediction coverage to include human disease.

IMP identifies homologs with conserved functional roles for disease knowledge

transfer, allowing biologists to analyze disease contexts and predictions across

all organisms. Additionally, IMP 2.0 implements a new flexible platform for

experts to generate custom hypotheses about biological processes or diseases,

making sophisticated data-driven methods easily accessible to researchers. IMP

does not require any registration or installation and is freely available for use

at http://imp.princeton.edu.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv486

PMCID: PMC4489318

PMID: 25969450 [Indexed for MEDLINE]

497. Nucleic Acids Res. 2015 Jul 1;43(W1):W431-5. doi: 10.1093/nar/gkv495. Epub 2015

May 12.

GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and

energy optimization.

Lee H(1), Heo L(1), Lee MS(2), Seok C(3).

Author information:

(1)Department of Chemistry, Seoul National University, Seoul 151-747, Korea.

(2)Department of Biomedical Sciences, College of Medicine, University of Ulsan,

Seoul 138-736, Korea chaok@snu.ac.kr. (3)Department of Chemistry, Seoul National

University, Seoul 151-747, Korea chaok@snu.ac.kr.

Protein-peptide interactions are involved in a wide range of biological processes

and are attractive targets for therapeutic purposes because of their small

interfaces. Therefore, effective protein-peptide docking techniques can provide

the basis for potential therapeutic applications by enabling an atomic-level

understanding of protein interactions. With the increasing number of

protein-peptide structures deposited in the protein data bank, the prediction

accuracy of protein-peptide docking can be enhanced by utilizing the information

provided by the database. The GalaxyPepDock web server, which is freely

accessible at http://galaxy.seoklab.org/pepdock, performs similarity-based

docking by finding templates from the database of experimentally determined

structures and building models using energy-based optimization that allows for

structural flexibility. The server can therefore effectively model the structural

differences between the template and target protein-peptide complexes. The

performance of GalaxyPepDock is superior to those of the other currently

available web servers when tested on the PeptiDB set and on recently released

complex structures. When tested on the CAPRI target 67, GalaxyPepDock generates

models that are more accurate than the best server models submitted during the

CAPRI blind prediction experiment.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv495

PMCID: PMC4489314

PMID: 25969449 [Indexed for MEDLINE]

498. Nucleic Acids Res. 2015 Jul 1;43(W1):W566-70. doi: 10.1093/nar/gkv468. Epub 2015

May 12.

ClustVis: a web tool for visualizing clustering of multivariate data using

Principal Component Analysis and heatmap.

Metsalu T(1), Vilo J(2).

Author information:

(1)Institute of Computer Science, University of Tartu, J. Liivi 2, 50409, Tartu,

Estonia. (2)Institute of Computer Science, University of Tartu, J. Liivi 2,

50409, Tartu, Estonia vilo@ut.ee.

The Principal Component Analysis (PCA) is a widely used method of reducing the

dimensionality of high-dimensional data, often followed by visualizing two of the

components on the scatterplot. Although widely used, the method is lacking an

easy-to-use web interface that scientists with little programming skills could

use to make plots of their own data. The same applies to creating heatmaps: it is

possible to add conditional formatting for Excel cells to show colored heatmaps,

but for more advanced features such as clustering and experimental annotations,

more sophisticated analysis tools have to be used. We present a web tool called

ClustVis that aims to have an intuitive user interface. Users can upload data

from a simple delimited text file that can be created in a spreadsheet program.

It is possible to modify data processing methods and the final appearance of the

PCA and heatmap plots by using drop-down menus, text boxes, sliders etc.

Appropriate defaults are given to reduce the time needed by the user to specify

input parameters. As an output, users can download PCA plot and heatmap in one of

the preferred file formats. This web server is freely available at

http://biit.cs.ut.ee/clustvis/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv468

PMCID: PMC4489295

PMID: 25969447 [Indexed for MEDLINE]

499. Nucleic Acids Res. 2015 Jul 1;43(W1):W401-7. doi: 10.1093/nar/gkv485. Epub 2015

May 12.

The TOPCONS web server for consensus prediction of membrane protein topology and

signal peptides.

Tsirigos KD(1), Peters C(1), Shu N(2), Käll L(1), Elofsson A(3).

Author information:

(1)Department of Biochemistry and Biophysics, Stockholm University, 10691

Stockholm, Sweden Science for Life Laboratory, Stockholm University, Box 1031,

17121 Solna, Sweden. (2)Department of Biochemistry and Biophysics, Stockholm

University, 10691 Stockholm, Sweden Science for Life Laboratory, Stockholm

University, Box 1031, 17121 Solna, Sweden Bioinformatics Infrastructure for Life

Sciences (BILS), Stockholm University, Sweden. (3)Department of Biochemistry and

Biophysics, Stockholm University, 10691 Stockholm, Sweden Science for Life

Laboratory, Stockholm University, Box 1031, 17121 Solna, Sweden arne@bioinfo.se.

TOPCONS (http://topcons.net/) is a widely used web server for consensus

prediction of membrane protein topology. We hereby present a major update to the

server, with some substantial improvements, including the following: (i) TOPCONS

can now efficiently separate signal peptides from transmembrane regions. (ii) The

server can now differentiate more successfully between globular and membrane

proteins. (iii) The server now is even slightly faster, although a much larger

database is used to generate the multiple sequence alignments. For most proteins,

the final prediction is produced in a matter of seconds. (iv) The user-friendly

interface is retained, with the additional feature of submitting batch files and

accessing the server programmatically using standard interfaces, making it thus

ideal for proteome-wide analyses. Indicatively, the user can now scan the entire

human proteome in a few days. (v) For proteins with homology to a known 3D

structure, the homology-inferred topology is also displayed. (vi) Finally, the

combination of methods currently implemented achieves an overall increase in

performance by 4% as compared to the currently available best-scoring methods and

TOPCONS is the only method that can identify signal peptides and still maintain a

state-of-the-art performance in topology predictions.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv485

PMCID: PMC4489233

PMID: 25969446 [Indexed for MEDLINE]

500. Nucleic Acids Res. 2015 Jul 1;43(W1):W78-84. doi: 10.1093/nar/gkv487. Epub 2015

May 11.

OrthoVenn: a web server for genome wide comparison and annotation of orthologous

clusters across multiple species.

Wang Y(1), Coleman-Derr D(2), Chen G(3), Gu YQ(4).

Author information:

(1)USDA-ARS, Western Regional Research Center, Crop Improvement and Genetics

Research Unit, Albany, CA 94710, USA Department of Plant Sciences, University of

California, Davis, CA 95616, USA Bioengineering College, Campus A, Chongqing

University, Chongqing 400030, China Yi.Wang@ars.usda.gov. (2)USDA-ARS, Plant Gene

Expression Center, Albany, CA 94710, USA. (3)Bioengineering College, Campus A,

Chongqing University, Chongqing 400030, China. (4)USDA-ARS, Western Regional

Research Center, Crop Improvement and Genetics Research Unit, Albany, CA 94710,

USA Yi.Wang@ars.usda.gov.

Genome wide analysis of orthologous clusters is an important component of

comparative genomics studies. Identifying the overlap among orthologous clusters

can enable us to elucidate the function and evolution of proteins across multiple

species. Here, we report a web platform named OrthoVenn that is useful for genome

wide comparisons and visualization of orthologous clusters. OrthoVenn provides

coverage of vertebrates, metazoa, protists, fungi, plants and bacteria for the

comparison of orthologous clusters and also supports uploading of customized

protein sequences from user-defined species. An interactive Venn diagram, summary

counts, and functional summaries of the disjunction and intersection of clusters

shared between species are displayed as part of the OrthoVenn result. OrthoVenn

also includes in-depth views of the clusters using various sequence analysis

tools. Furthermore, OrthoVenn identifies orthologous clusters of single copy

genes and allows for a customized search of clusters of specific genes through

key words or BLAST. OrthoVenn is an efficient and user-friendly web server freely

accessible at http://probes.pw.usda.gov/OrthoVenn or

http://aegilops.wheat.ucdavis.edu/OrthoVenn.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv487

PMCID: PMC4489293

PMID: 25964301 [Indexed for MEDLINE]

501. Nucleic Acids Res. 2015 Jul 1;43(W1):W148-53. doi: 10.1093/nar/gkv488. Epub 2015

May 11.

CATH FunFHMMer web server: protein functional annotations using functional family

assignments.

Das S(1), Sillitoe I(2), Lee D(3), Lees JG(3), Dawson NL(3), Ward J(4), Orengo

CA(5).

Author information:

(1)Institute of Structural and Molecular Biology, UCL, Darwin Building, Gower

Street, WC1E 6BT, UK sayoni.das.12@ucl.ac.uk. (2)Institute of Structural and

Molecular Biology, UCL, Darwin Building, Gower Street, WC1E 6BT, UK

i.sillitoe@ucl.a.uk. (3)Institute of Structural and Molecular Biology, UCL,

Darwin Building, Gower Street, WC1E 6BT, UK. (4)Department of Biochemical

Engineering, UCL, Gower Street, WC1E 6BT, UK. (5)Institute of Structural and

Molecular Biology, UCL, Darwin Building, Gower Street, WC1E 6BT, UK

c.orengo@ucl.ac.uk.

The widening function annotation gap in protein databases and the increasing

number and diversity of the proteins being sequenced presents new challenges to

protein function prediction methods. Multidomain proteins complicate the protein

sequence-structure-function relationship further as new combinations of domains

can expand the functional repertoire, creating new proteins and functions. Here,

we present the FunFHMMer web server, which provides Gene Ontology (GO)

annotations for query protein sequences based on the functional classification of

the domain-based CATH-Gene3D resource. Our server also provides valuable

information for the prediction of functional sites. The predictive power of

FunFHMMer has been validated on a set of 95 proteins where FunFHMMer performs

better than BLAST, Pfam and CDD. Recent validation by an independent

international competition ranks FunFHMMer as one of the top function prediction

methods in predicting GO annotations for both the Biological Process and

Molecular Function Ontology. The FunFHMMer web server is available at

http://www.cathdb.info/search/by\_funfhmmer.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv488

PMCID: PMC4489299

PMID: 25964299 [Indexed for MEDLINE]

502. Nucleic Acids Res. 2015 Jul 1;43(W1):W498-501. doi: 10.1093/nar/gkv465. Epub 2015

May 11.

RNA-Redesign: a web server for fixed-backbone 3D design of RNA.

Yesselman JD(1), Das R(2).

Author information:

(1)Biochemistry Department, Stanford University, Stanford, CA 94305, USA.

(2)Biochemistry Department, Stanford University, Stanford, CA 94305, USA Physics

Department, Stanford University, Stanford, CA 94305, USA rhiju@stanford.edu.

RNA is rising in importance as a design medium for interrogating fundamental

biology and for developing therapeutic and bioengineering applications. While

there are several online servers for design of RNA secondary structure, there are

no tools available for the rational design of 3D RNA structure. Here we present

RNA-Redesign (http://rnaredesign.stanford.edu), an online 3D design tool for RNA.

This resource utilizes fixed-backbone design to optimize the sequence identity

and nucleobase conformations of an RNA to match a desired backbone, analogous to

fundamental tools that underlie rational protein engineering. The resulting

sequences suggest thermostabilizing mutations that can be experimentally

verified. Further, sequence preferences that differ between natural and

computationally designed sequences can suggest whether natural sequences possess

functional constraints besides folding stability, such as cofactor binding or

conformational switching. Finally, for biochemical studies, the designed

sequences can suggest experimental tests of 3D models, including concomitant

mutation of base triples. In addition to the designs generated, detailed

graphical analysis is presented through an integrated and user-friendly

environment.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv465

PMCID: PMC4489241

PMID: 25964298 [Indexed for MEDLINE]

503. Nucleic Acids Res. 2015 Jul 1;43(W1):W65-71. doi: 10.1093/nar/gkv458. Epub 2015

May 9.

Pse-in-One: a web server for generating various modes of pseudo components of

DNA, RNA, and protein sequences.

Liu B(1), Liu F(2), Wang X(3), Chen J(2), Fang L(2), Chou KC(4).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong 518055, China Key Laboratory of

Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, Guangdong 518055, China Gordon Life Science Institute,

Belmont, MA 02478, USA bliu@gordonlifescience.org bliu@insun.hit.edu.cn.

(2)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong 518055, China. (3)School of

Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate

School, Shenzhen, Guangdong 518055, China Key Laboratory of Network Oriented

Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School,

Shenzhen, Guangdong 518055, China. (4)Gordon Life Science Institute, Belmont, MA

02478, USA Center of Excellence in Genomic Medicine Research (CEGMR), King

Abdulaziz University, Jeddah 21589, Saudi Arabia kcchou@gordonlifescience.org.

With the avalanche of biological sequences generated in the post-genomic age, one

of the most challenging problems in computational biology is how to effectively

formulate the sequence of a biological sample (such as DNA, RNA or protein) with

a discrete model or a vector that can effectively reflect its sequence pattern

information or capture its key features concerned. Although several web servers

and stand-alone tools were developed to address this problem, all these tools,

however, can only handle one type of samples. Furthermore, the number of their

built-in properties is limited, and hence it is often difficult for users to

formulate the biological sequences according to their desired features or

properties. In this article, with a much larger number of built-in properties, we

are to propose a much more flexible web server called Pse-in-One

(http://bioinformatics.hitsz.edu.cn/Pse-in-One/), which can, through its 28

different modes, generate nearly all the possible feature vectors for DNA, RNA

and protein sequences. Particularly, it can also generate those feature vectors

with the properties defined by users themselves. These feature vectors can be

easily combined with machine-learning algorithms to develop computational

predictors and analysis methods for various tasks in bioinformatics and system

biology. It is anticipated that the Pse-in-One web server will become a very

useful tool in computational proteomics, genomics, as well as biological sequence

analysis. Moreover, to maximize users' convenience, its stand-alone version can

also be downloaded from http://bioinformatics.hitsz.edu.cn/Pse-in-One/download/,

and directly run on Windows, Linux, Unix and Mac OS.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv458

PMCID: PMC4489303

PMID: 25958395 [Indexed for MEDLINE]

504. Nucleic Acids Res. 2015 Jul 1;43(W1):W560-5. doi: 10.1093/nar/gkv450. Epub 2015

May 9.

NaviCell Web Service for network-based data visualization.

Bonnet E(1), Viara E(2), Kuperstein I(1), Calzone L(1), Cohen DP(1), Barillot

E(1), Zinovyev A(3).

Author information:

(1)Institut Curie, 26 rue d'Ulm, 75248 Paris, France INSERM, U900, 75248 Paris,

France Mines ParisTech, 77300 Fontainebleau, France. (2)Sysra, 91330 Yerres,

France. (3)Institut Curie, 26 rue d'Ulm, 75248 Paris, France INSERM, U900, 75248

Paris, France Mines ParisTech, 77300 Fontainebleau, France

andrei.zinovyev@curie.fr.

Data visualization is an essential element of biological research, required for

obtaining insights and formulating new hypotheses on mechanisms of health and

disease. NaviCell Web Service is a tool for network-based visualization of

'omics' data which implements several data visual representation methods and

utilities for combining them together. NaviCell Web Service uses Google Maps and

semantic zooming to browse large biological network maps, represented in various

formats, together with different types of the molecular data mapped on top of

them. For achieving this, the tool provides standard heatmaps, barplots and

glyphs as well as the novel map staining technique for grasping large-scale

trends in numerical values (such as whole transcriptome) projected onto a pathway

map. The web service provides a server mode, which allows automating

visualization tasks and retrieving data from maps via RESTful (standard HTTP)

calls. Bindings to different programming languages are provided (Python and R).

We illustrate the purpose of the tool with several case studies using pathway

maps created by different research groups, in which data visualization provides

new insights into molecular mechanisms involved in systemic diseases such as

cancer and neurodegenerative diseases.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv450

PMCID: PMC4489283

PMID: 25958393 [Indexed for MEDLINE]

505. Nucleic Acids Res. 2015 Jul 1;43(W1):W362-9. doi: 10.1093/nar/gkv463. Epub 2015

May 8.

xVis: a web server for the schematic visualization and interpretation of

crosslink-derived spatial restraints.

Grimm M(1), Zimniak T(1), Kahraman A(2), Herzog F(3).

Author information:

(1)Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität

München, Munich 81377, Germany. (2)Institute of Molecular Life Sciences,

University of Zurich, Zurich 8057, Switzerland. (3)Gene Center and Department of

Biochemistry, Ludwig-Maximilians-Universität München, Munich 81377, Germany

herzog@genzentrum.lmu.de.

The identification of crosslinks by mass spectrometry has recently been

established as an integral part of the hybrid structural analysis of protein

complexes and networks. The crosslinking analysis determines distance restraints

between two covalently linked amino acids which are typically summarized in a

table format that precludes the immediate and comprehensive interpretation of the

topological data. xVis displays crosslinks in clear schematic representations in

form of a circular, bar or network diagram. The interactive graphs indicate the

linkage sites and identification scores, depict the spatial proximity of

structurally and functionally annotated protein regions and the evolutionary

conservation of amino acids and facilitate clustering of proteins into

subcomplexes according to the crosslink density. Furthermore, xVis offers two

options for the qualitative assessment of the crosslink identifications by

filtering crosslinks according to identification scores or false discovery rates

and by displaying the corresponding fragment ion spectrum of each crosslink for

the manual validation of the mass spectrometric data. Our web server provides an

easy-to-use tool for the fast topological and functional interpretation of

distance information on protein complex architectures and for the evaluation of

crosslink fragment ion spectra. xVis is available under a Creative Commons

Attribution-ShareAlike 4.0 International license at

http://xvis.genzentrum.lmu.de/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv463

PMCID: PMC4489277

PMID: 25956653 [Indexed for MEDLINE]

506. Nucleic Acids Res. 2015 Jul 1;43(W1):W436-42. doi: 10.1093/nar/gkv462. Epub 2015

May 8.

PockDrug-Server: a new web server for predicting pocket druggability on holo and

apo proteins.

Hussein HA(1), Borrel A(2), Geneix C(1), Petitjean M(1), Regad L(1), Camproux

AC(3).

Author information:

(1)INSERM, UMRS-973, MTi, Université Paris Diderot, 35 Rue Hélène Brion, 75205

Paris Cedex 13, case courier 7113, Paris, France Université Paris Diderot,

Sorbonne Paris Cité, UMRS-973, MTi, Paris, France. (2)INSERM, UMRS-973, MTi,

Université Paris Diderot, 35 Rue Hélène Brion, 75205 Paris Cedex 13, case courier

7113, Paris, France Université Paris Diderot, Sorbonne Paris Cité, UMRS-973, MTi,

Paris, France Division of Pharmaceutical Chemistry, Faculty of pharmacy,

University of Helsinki, Viikinkaari 9 (P.O. Box 56) FI-00014, Finland. (3)INSERM,

UMRS-973, MTi, Université Paris Diderot, 35 Rue Hélène Brion, 75205 Paris Cedex

13, case courier 7113, Paris, France Université Paris Diderot, Sorbonne Paris

Cité, UMRS-973, MTi, Paris, France anne-claude.camproux@univ-paris-diderot.fr.

Predicting protein pocket's ability to bind drug-like molecules with high

affinity, i.e. druggability, is of major interest in the target identification

phase of drug discovery. Therefore, pocket druggability investigations represent

a key step of compound clinical progression projects. Currently computational

druggability prediction models are attached to one unique pocket estimation

method despite pocket estimation uncertainties. In this paper, we propose

'PockDrug-Server' to predict pocket druggability, efficient on both (i) estimated

pockets guided by the ligand proximity (extracted by proximity to a ligand from a

holo protein structure) and (ii) estimated pockets based solely on protein

structure information (based on amino atoms that form the surface of potential

binding cavities). PockDrug-Server provides consistent druggability results using

different pocket estimation methods. It is robust with respect to pocket boundary

and estimation uncertainties, thus efficient using apo pockets that are

challenging to estimate. It clearly distinguishes druggable from less druggable

pockets using different estimation methods and outperformed recent druggability

models for apo pockets. It can be carried out from one or a set of apo/holo

proteins using different pocket estimation methods proposed by our web server or

from any pocket previously estimated by the user. PockDrug-Server is publicly

available at: http://pockdrug.rpbs.univ-paris-diderot.fr.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv462

PMCID: PMC4489252

PMID: 25956651 [Indexed for MEDLINE]

507. Nucleic Acids Res. 2015 Jul 1;43(W1):W39-49. doi: 10.1093/nar/gkv416. Epub 2015

May 7.

The MEME Suite.

Bailey TL(1), Johnson J(2), Grant CE(3), Noble WS(4).

Author information:

(1)Institute for Molecular Bioscience, The University of Queensland, Brisbane

4072, Queensland, Australia t.bailey@uq.edu.au. (2)Institute for Molecular

Bioscience, The University of Queensland, Brisbane 4072, Queensland, Australia.

(3)Department of Genome Sciences, Department of Computer Science and Engineering,

University of Washington, Seattle, WA 98109, USA. (4)Department of Genome

Sciences, Department of Computer Science and Engineering, University of

Washington, Seattle, WA 98109, USA william-noble@uw.edu.

The MEME Suite is a powerful, integrated set of web-based tools for studying

sequence motifs in proteins, DNA and RNA. Such motifs encode many biological

functions, and their detection and characterization is important in the study of

molecular interactions in the cell, including the regulation of gene expression.

Since the previous description of the MEME Suite in the 2009 Nucleic Acids

Research Web Server Issue, we have added six new tools. Here we describe the

capabilities of all the tools within the suite, give advice on their best use and

provide several case studies to illustrate how to combine the results of various

MEME Suite tools for successful motif-based analyses. The MEME Suite is freely

available for academic use at http://meme-suite.org, and source code is also

available for download and local installation.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv416

PMCID: PMC4489269

PMID: 25953851 [Indexed for MEDLINE]

508. Nucleic Acids Res. 2015 Jul 1;43(W1):W474-9. doi: 10.1093/nar/gkv457. Epub 2015

May 6.

PACCMIT/PACCMIT-CDS: identifying microRNA targets in 3' UTRs and coding

sequences.

Šulc M(1), Marín RM(1), Robins HS(2), Vaníček J(3).

Author information:

(1)Laboratory of Theoretical Physical Chemistry, Institut des Sciences et

Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015

Lausanne, Switzerland. (2)Fred Hutchinson Cancer Research Center, Seattle, WA

98109, USA. (3)Laboratory of Theoretical Physical Chemistry, Institut des

Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne

(EPFL), CH-1015 Lausanne, Switzerland jiri.vanicek@epfl.ch.

The purpose of the proposed web server, publicly available at

http://paccmit.epfl.ch, is to provide a user-friendly interface to two algorithms

for predicting messenger RNA (mRNA) molecules regulated by microRNAs: (i) PACCMIT

(Prediction of ACcessible and/or Conserved MIcroRNA Targets), which identifies

primarily mRNA transcripts targeted in their 3' untranslated regions (3' UTRs),

and (ii) PACCMIT-CDS, designed to find mRNAs targeted within their coding

sequences (CDSs). While PACCMIT belongs among the accurate algorithms for

predicting conserved microRNA targets in the 3' UTRs, the main contribution of

the web server is 2-fold: PACCMIT provides an accurate tool for predicting

targets also of weakly conserved or non-conserved microRNAs, whereas PACCMIT-CDS

addresses the lack of similar portals adapted specifically for targets in CDS.

The web server asks the user for microRNAs and mRNAs to be analyzed, accesses the

precomputed P-values for all microRNA-mRNA pairs from a database for all mRNAs

and microRNAs in a given species, ranks the predicted microRNA-mRNA pairs,

evaluates their significance according to the false discovery rate and finally

displays the predictions in a tabular form. The results are also available for

download in several standard formats.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv457

PMCID: PMC4489287

PMID: 25948580 [Indexed for MEDLINE]

509. Nucleic Acids Res. 2015 Jul 1;43(W1):W237-43. doi: 10.1093/nar/gkv437. Epub 2015

May 6.

antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene

clusters.

Weber T(1), Blin K(2), Duddela S(3), Krug D(4), Kim HU(5), Bruccoleri R(6), Lee

SY(5), Fischbach MA(7), Müller R(4), Wohlleben W(8), Breitling R(9), Takano E(9),

Medema MH(10).

Author information:

(1)Novo Nordisk Foundation Center for Biosustainability, Technical University of

Denmark, Hørsholm, Denmark marnix.medema@wur.nl. (2)Novo Nordisk Foundation

Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark.

(3)Department of Microbial Natural Products, Helmholtz Institute for

Pharmaceutical Research Saarland, Helmholtz Centre for Infection Research,

Saarland University, Saarbrücken, Germany. (4)Department of Microbial Natural

Products, Helmholtz Institute for Pharmaceutical Research Saarland, Helmholtz

Centre for Infection Research, Saarland University, Saarbrücken, Germany German

Centre for Infection Research (DZIF), Location Hannover-Braunschweig, Germany.

(5)Novo Nordisk Foundation Center for Biosustainability, Technical University of

Denmark, Hørsholm, Denmark Department of Chemical and Biomolecular Engineering

(BK21 Plus Program) / BioInformatics Research Center, Korea Advanced Institute of

Science and Technology, Daejeon, Republic of Korea. (6)Congenomics, LLC,

Glastonbury, CT, USA. (7)Department of Bioengineering and Therapeutic

Sciences/California Institute for Quantitative Biosciences, University of

California, San Francisco, USA. (8)Interfaculty Institute for Microbiology and

Infection Medicine, Eberhard Karls University of Tübingen, Tübingen, Germany

German Center for Infection Research (DZIF), Location Tübingen, Germany.

(9)Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals

(SYNBIOCHEM), Manchester Institute of Biotechnology, Faculty of Life Sciences,

The University of Manchester, Manchester, UK. (10)Microbial Genomics and

Bioinformatics Research Group, Max Planck Institute for Marine Microbiology,

Bremen, Germany Bioinformatics Group, Wageningen University, Wageningen, The

Netherlands marnix.medema@wur.nl.

Microbial secondary metabolism constitutes a rich source of antibiotics,

chemotherapeutics, insecticides and other high-value chemicals. Genome mining of

gene clusters that encode the biosynthetic pathways for these metabolites has

become a key methodology for novel compound discovery. In 2011, we introduced

antiSMASH, a web server and stand-alone tool for the automatic genomic

identification and analysis of biosynthetic gene clusters, available at

http://antismash.secondarymetabolites.org. Here, we present version 3.0 of

antiSMASH, which has undergone major improvements. A full integration of the

recently published ClusterFinder algorithm now allows using this probabilistic

algorithm to detect putative gene clusters of unknown types. Also, a new

dereplication variant of the ClusterBlast module now identifies similarities of

identified clusters to any of 1172 clusters with known end products. At the

enzyme level, active sites of key biosynthetic enzymes are now pinpointed through

a curated pattern-matching procedure and Enzyme Commission numbers are assigned

to functionally classify all enzyme-coding genes. Additionally, chemical

structure prediction has been improved by incorporating polyketide reduction

states. Finally, in order for users to be able to organize and analyze multiple

antiSMASH outputs in a private setting, a new XML output module allows offline

editing of antiSMASH annotations within the Geneious software.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv437

PMCID: PMC4489286

PMID: 25948579 [Indexed for MEDLINE]

510. Nucleic Acids Res. 2015 Jul 1;43(W1):W408-12. doi: 10.1093/nar/gkv451. Epub 2015

May 5.

CCTOP: a Consensus Constrained TOPology prediction web server.

Dobson L(1), Reményi I(1), Tusnády GE(2).

Author information:

(1)'Momentum' Membrane Protein Bioinformatics Research Group, Institute of

Enzymology, RCNS, HAS, Budapest PO Box 7, H-1518, Hungary. (2)'Momentum' Membrane

Protein Bioinformatics Research Group, Institute of Enzymology, RCNS, HAS,

Budapest PO Box 7, H-1518, Hungary tusnady.gabor@ttk.mta.hu.

The Consensus Constrained TOPology prediction (CCTOP;

http://cctop.enzim.ttk.mta.hu) server is a web-based application providing

transmembrane topology prediction. In addition to utilizing 10 different

state-of-the-art topology prediction methods, the CCTOP server incorporates

topology information from existing experimental and computational sources

available in the PDBTM, TOPDB and TOPDOM databases using the probabilistic

framework of hidden Markov model. The server provides the option to precede the

topology prediction with signal peptide prediction and transmembrane-globular

protein discrimination. The initial result can be recalculated by (de)selecting

any of the prediction methods or mapped experiments or by adding user specified

constraints. CCTOP showed superior performance to existing approaches. The

reliability of each prediction is also calculated, which correlates with the

accuracy of the per protein topology prediction. The prediction results and the

collected experimental information are visualized on the CCTOP home page and can

be downloaded in XML format. Programmable access of the CCTOP server is also

available, and an example of client-side script is provided.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv451

PMCID: PMC4489262

PMID: 25943549 [Indexed for MEDLINE]

511. Nucleic Acids Res. 2015 Jul 1;43(W1):W30-8. doi: 10.1093/nar/gkv397. Epub 2015

May 5.

HMMER web server: 2015 update.

Finn RD(1), Clements J(2), Arndt W(2), Miller BL(2), Wheeler TJ(3), Schreiber

F(4), Bateman A(4), Eddy SR(2).

Author information:

(1)European Molecular Biology Laboratory, European Bioinformatics Institute

(EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK HHMI

Janelia Research Campus, 19700 Helix Drive, Ashburn, VA 20147, USA rdf@ebi.ac.uk.

(2)HHMI Janelia Research Campus, 19700 Helix Drive, Ashburn, VA 20147, USA.

(3)HHMI Janelia Research Campus, 19700 Helix Drive, Ashburn, VA 20147, USA

Department of Computer Science, University of Montana, Social Sciences Building

Room 412, Missoula MT 59812, USA. (4)European Molecular Biology Laboratory,

European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus,

Hinxton, Cambridge, CB10 1SD, UK.

The HMMER website, available at http://www.ebi.ac.uk/Tools/hmmer/, provides

access to the protein homology search algorithms found in the HMMER software

suite. Since the first release of the website in 2011, the search repertoire has

been expanded to include the iterative search algorithm, jackhmmer. The continued

growth of the target sequence databases means that traditional tabular

representations of significant sequence hits can be overwhelming to the user.

Consequently, additional ways of presenting homology search results have been

developed, allowing them to be summarised according to taxonomic distribution or

domain architecture. The taxonomy and domain architecture representations can be

used in combination to filter the results according to the needs of a user.

Searches can also be restricted prior to submission using a new taxonomic filter,

which not only ensures that the results are specific to the requested taxonomic

group, but also improves search performance. The repertoire of profile hidden

Markov model libraries, which are used for annotation of query sequences with

protein families and domains, has been expanded to include the libraries from

CATH-Gene3D, PIRSF, Superfamily and TIGRFAMs. Finally, we discuss the relocation

of the HMMER webserver to the European Bioinformatics Institute and the potential

impact that this will have.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv397

PMCID: PMC4489315

PMID: 25943547 [Indexed for MEDLINE]

512. Nucleic Acids Res. 2015 Jul 1;43(W1):W338-42. doi: 10.1093/nar/gkv454. Epub 2015

May 5.

(PS)2: protein structure prediction server version 3.0.

Huang TT(1), Hwang JK(2), Chen CH(3), Chu CS(4), Lee CW(5), Chen CC(6).

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu 30068, Taiwan Center for Bioinformatics Research, National

Chiao Tung University, Hsinchu 30068, Taiwan. (2)Institute of Bioinformatics and

Systems Biology, National Chiao Tung University, Hsinchu 30068, Taiwan Center for

Bioinformatics Research, National Chiao Tung University, Hsinchu 30068, Taiwan

Department of Biological Science and Technology, National Chiao Tung University,

Hsinchu 30068, Taiwan Department of Bioinformatics and Medical Engineering, Asia

University, Taichung 41354, Taiwan. (3)Department of Vascular and Medicinal

Research, Texas Heart Institute, Houston, TX 77030, USA Center for Lipid

Biosciences, Kaohsiung Medical University Hospital, Kaohsiung Medical University,

Kaohsiung 80708, Taiwan L5 Research Center, China Medical University Hospital,

China Medical University, Taichung 40402, Taiwan Section of Cardiovascular

Research, Department of Medicine, Baylor College of Medicine, Houston, TX 77030,

USA. (4)Center for Lipid Biosciences, Kaohsiung Medical University Hospital,

Kaohsiung Medical University, Kaohsiung 80708, Taiwan Department of Internal

Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University,

Kaohsiung 80708, Taiwan. (5)Center for Bioinformatics Research, National Chiao

Tung University, Hsinchu 30068, Taiwan Department of Biological Science and

Technology, National Chiao Tung University, Hsinchu 30068, Taiwan. (6)Center for

Lipid Biosciences, Kaohsiung Medical University Hospital, Kaohsiung Medical

University, Kaohsiung 80708, Taiwan Center for Lipid and Glycomedicine Research,

Kaohsiung Medical University, Kaohsiung 80708, Taiwan Institute of Medical

Science and Technology, National Sun Yat-sen University, Kaohsiung 80424, Taiwan

chieh.bi91g@nctu.edu.tw.

Protein complexes are involved in many biological processes. Examining coupling

between subunits of a complex would be useful to understand the molecular basis

of protein function. Here, our updated (PS)(2) web server predicts the

three-dimensional structures of protein complexes based on comparative modeling;

furthermore, this server examines the coupling between subunits of the predicted

complex by combining structural and evolutionary considerations. The predicted

complex structure could be indicated and visualized by Java-based 3D graphics

viewers and the structural and evolutionary profiles are shown and compared

chain-by-chain. For each subunit, considerations with or without the packing

contribution of other subunits cause the differences in similarities between

structural and evolutionary profiles, and these differences imply which form,

complex or monomeric, is preferred in the biological condition for the subunit.

We believe that the (PS)(2) server would be a useful tool for biologists who are

interested not only in the structures of protein complexes but also in the

coupling between subunits of the complexes. The (PS)(2) is freely available at

http://ps2v3.life.nctu.edu.tw/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv454

PMCID: PMC4489310

PMID: 25943546 [Indexed for MEDLINE]

513. Nucleic Acids Res. 2015 Jul 1;43(W1):W419-24. doi: 10.1093/nar/gkv456. Epub 2015

May 5.

CABS-dock web server for the flexible docking of peptides to proteins without

prior knowledge of the binding site.

Kurcinski M(1), Jamroz M(1), Blaszczyk M(1), Kolinski A(2), Kmiecik S(3).

Author information:

(1)Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw,

Poland. (2)Department of Chemistry, University of Warsaw, Pasteura 1, 02-093

Warsaw, Poland kolinski@chem.uw.edu.pl. (3)Department of Chemistry, University of

Warsaw, Pasteura 1, 02-093 Warsaw, Poland sekmi@chem.uw.edu.pl.

Protein-peptide interactions play a key role in cell functions. Their structural

characterization, though challenging, is important for the discovery of new

drugs. The CABS-dock web server provides an interface for modeling

protein-peptide interactions using a highly efficient protocol for the flexible

docking of peptides to proteins. While other docking algorithms require

pre-defined localization of the binding site, CABS-dock does not require such

knowledge. Given a protein receptor structure and a peptide sequence (and

starting from random conformations and positions of the peptide), CABS-dock

performs simulation search for the binding site allowing for full flexibility of

the peptide and small fluctuations of the receptor backbone. This protocol was

extensively tested over the largest dataset of non-redundant protein-peptide

interactions available to date (including bound and unbound docking cases). For

over 80% of bound and unbound dataset cases, we obtained models with high or

medium accuracy (sufficient for practical applications). Additionally, as

optional features, CABS-dock can exclude user-selected binding modes from docking

search or to increase the level of flexibility for chosen receptor fragments.

CABS-dock is freely available as a web server at

http://biocomp.chem.uw.edu.pl/CABSdock.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv456

PMCID: PMC4489223

PMID: 25943545 [Indexed for MEDLINE]

514. Nucleic Acids Res. 2015 Jul 1;43(W1):W91-7. doi: 10.1093/nar/gkv453. Epub 2015

May 5.

FlyNet: a versatile network prioritization server for the Drosophila community.

Shin J(1), Yang S(1), Kim E(1), Kim CY(1), Shim H(1), Cho A(1), Kim H(1), Hwang

S(1), Shim JE(1), Lee I(2).

Author information:

(1)Department of Biotechnology, College of Life Science and Biotechnology, Yonsei

University, Seoul, Korea. (2)Department of Biotechnology, College of Life Science

and Biotechnology, Yonsei University, Seoul, Korea insuklee@yonsei.ac.kr.

Drosophila melanogaster (fruit fly) has been a popular model organism in animal

genetics due to the high accessibility of reverse-genetics tools. In addition,

the close relationship between the Drosophila and human genomes rationalizes the

use of Drosophila as an invertebrate model for human neurobiology and disease

research. A platform technology for predicting candidate genes or functions would

further enhance the usefulness of this long-established model organism for

gene-to-phenotype mapping. Recently, the power of network prioritization for

gene-to-phenotype mapping has been demonstrated in many organisms. Here we

present a network prioritization server dedicated to Drosophila that covers ∼95%

of the coding genome. This server, dubbed FlyNet, has several distinctive

features, including (i) prioritization for both genes and functions; (ii) two

complementary network algorithms: direct neighborhood and network diffusion;

(iii) spatiotemporal-specific networks as an additional prioritization strategy

for traits associated with a specific developmental stage or tissue and (iv)

prioritization for human disease genes. FlyNet is expected to serve as a

versatile hypothesis-generation platform for genes and functions in the study of

basic animal genetics, developmental biology and human disease. FlyNet is

available for free at http://www.inetbio.org/flynet.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv453

PMCID: PMC4489278

PMID: 25943544 [Indexed for MEDLINE]

515. Nucleic Acids Res. 2015 Jul 1;43(W1):W283-8. doi: 10.1093/nar/gkv418. Epub 2015

May 5.

TFmiR: a web server for constructing and analyzing disease-specific transcription

factor and miRNA co-regulatory networks.

Hamed M(1), Spaniol C(1), Nazarieh M(1), Helms V(2).

Author information:

(1)Center for Bioinformatics, Saarland University, 66041 Saarbrucken, Germany.

(2)Center for Bioinformatics, Saarland University, 66041 Saarbrucken, Germany

volkhard.helms@bioinformatik.uni-saarland.de.

TFmiR is a freely available web server for deep and integrative analysis of

combinatorial regulatory interactions between transcription factors, microRNAs

and target genes that are involved in disease pathogenesis. Since the inner

workings of cells rely on the correct functioning of an enormously complex system

of activating and repressing interactions that can be perturbed in many ways,

TFmiR helps to better elucidate cellular mechanisms at the molecular level from a

network perspective. The provided topological and functional analyses promote

TFmiR as a reliable systems biology tool for researchers across the life science

communities. TFmiR web server is accessible through the following URL:

http://service.bioinformatik.uni-saarland.de/tfmir.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv418

PMCID: PMC4489273

PMID: 25943543 [Indexed for MEDLINE]

516. Nucleic Acids Res. 2015 Jul 1;43(W1):W182-7. doi: 10.1093/nar/gkv443. Epub 2015

May 4.

FNTM: a server for predicting functional networks of tissues in mouse.

Goya J(1), Wong AK(2), Yao V(3), Krishnan A(1), Homilius M(3), Troyanskaya OG(4).

Author information:

(1)Lewis-Sigler Institute for Integrative Genomics, Princeton University,

Princeton, NJ 08540, USA. (2)Lewis-Sigler Institute for Integrative Genomics,

Princeton University, Princeton, NJ 08540, USA Simons Center for Data Analysis,

Simons Foundation, NY 10010, USA Department of Computer Science, Princeton

University, Princeton, NJ 08540, USA. (3)Lewis-Sigler Institute for Integrative

Genomics, Princeton University, Princeton, NJ 08540, USA Department of Computer

Science, Princeton University, Princeton, NJ 08540, USA. (4)Lewis-Sigler

Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540,

USA Simons Center for Data Analysis, Simons Foundation, NY 10010, USA Department

of Computer Science, Princeton University, Princeton, NJ 08540, USA

ogt@cs.princeton.edu.

Functional Networks of Tissues in Mouse (FNTM) provides biomedical researchers

with tissue-specific predictions of functional relationships between proteins in

the most widely used model organism for human disease, the laboratory mouse.

Users can explore FNTM-predicted functional relationships for their tissues and

genes of interest or examine gene function and interaction predictions across

multiple tissues, all through an interactive, multi-tissue network browser. FNTM

makes predictions based on integration of a variety of functional genomic data,

including over 13 000 gene expression experiments, and prior knowledge of gene

function. FNTM is an ideal starting point for clinical and translational

researchers considering a mouse model for their disease of interest, researchers

already working with mouse models who are interested in discovering new genes

related to their pathways or phenotypes of interest, and biologists working with

other organisms to explore the functional relationships of their genes of

interest in specific mouse tissue contexts. FNTM predicts tissue-specific

functional relationships in 200 tissues, does not require any registration or

installation and is freely available for use at http://fntm.princeton.edu.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv443

PMCID: PMC4489275

PMID: 25940632 [Indexed for MEDLINE]

517. Nucleic Acids Res. 2015 Jul 1;43(W1):W547-51. doi: 10.1093/nar/gkv417. Epub 2015

May 4.

MapMyFlu: visualizing spatio-temporal relationships between related influenza

sequences.

Nolte N(1), Kurzawa N(1), Eils R(1), Herrmann C(2).

Author information:

(1)Institute of Pharmacy and Molecular Biotechnology, and Bioquant Center,

University of Heidelberg, Im Neuenheimer Feld 267, Heidelberg 69120, Germany

Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Im

Neuenheimer Feld 580, Heidelberg 69120, Germany. (2)Institute of Pharmacy and

Molecular Biotechnology, and Bioquant Center, University of Heidelberg, Im

Neuenheimer Feld 267, Heidelberg 69120, Germany Division of Theoretical

Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580,

Heidelberg 69120, Germany carl.herrmann@uni-heidelberg.de.

Understanding the molecular dynamics of viral spreading is crucial for

anticipating the epidemiological implications of disease outbreaks. In the case

of influenza, reassortments or point mutations affect the adaption to new hosts

or resistance to anti-viral drugs and can determine whether a new strain will

result in a pandemic infection or a less severe progression. To this end, tools

integrating molecular information with epidemiological parameters are important

to understand how molecular characteristics reflect in the infection dynamics. We

present a new web tool, MapMyFlu, which allows to spatially and temporally

display influenza viruses related to a query sequence on a Google Map based on

BLAST results against the NCBI Influenza Database. Temporal and geographical

trends appear clearly and may help in reconstructing the evolutionary history of

a particular sequence. The tool is accessible through a web server, hence without

the need for local installation. The website has an intuitive design and provides

an easy-to-use service, and is available at

http://mapmyflu.ipmb.uni-heidelberg.de.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv417

PMCID: PMC4489300

PMID: 25940623 [Indexed for MEDLINE]

518. Nucleic Acids Res. 2015 Jul 1;43(W1):W507-12. doi: 10.1093/nar/gkv435. Epub 2015

May 4.

RNAPattMatch: a web server for RNA sequence/structure motif detection based on

pattern matching with flexible gaps.

Drory Retwitzer M(1), Polishchuk M(2), Churkin E(1), Kifer I(3), Yakhini Z(4),

Barash D(5).

Author information:

(1)Department of Computer Science, Ben-Gurion University, Beer-Sheva 84105,

Israel. (2)Department of Computer Science, Ben-Gurion University, Beer-Sheva

84105, Israel Department of Computational Biology, Vavilov Institute of General

Genetics, Russian Academy of Science, Moscow 11933, Russia. (3)Agilent

Laboratories, Agilent Technologies, Tel Aviv 49527, Israel Microsoft R&D Center,

Herzliya 46725 , Israel. (4)Agilent Laboratories, Agilent Technologies, Tel Aviv

49527, Israel Laboratory of Computational Biology, Computer Science Department,

Israel Institute of Technology, Haifa 32000, Israel. (5)Department of Computer

Science, Ben-Gurion University, Beer-Sheva 84105, Israel dbarash@cs.bgu.ac.il.

Searching for RNA sequence-structure patterns is becoming an essential tool for

RNA practitioners. Novel discoveries of regulatory non-coding RNAs in targeted

organisms and the motivation to find them across a wide range of organisms have

prompted the use of computational RNA pattern matching as an enhancement to

sequence similarity. State-of-the-art programs differ by the flexibility of

patterns allowed as queries and by their simplicity of use. In particular-no

existing method is available as a user-friendly web server. A general program

that searches for RNA sequence-structure patterns is RNA Structator. However, it

is not available as a web server and does not provide the option to allow

flexible gap pattern representation with an upper bound of the gap length being

specified at any position in the sequence. Here, we introduce RNAPattMatch, a

web-based application that is user friendly and makes sequence/structure RNA

queries accessible to practitioners of various background and proficiency. It

also extends RNA Structator and allows a more flexible variable gaps

representation, in addition to analysis of results using energy minimization

methods. RNAPattMatch service is available at

http://www.cs.bgu.ac.il/rnapattmatch. A standalone version of the search tool is

also available to download at the site.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv435

PMCID: PMC4489251

PMID: 25940619 [Indexed for MEDLINE]

519. Nucleic Acids Res. 2015 Jul 1;43(W1):W231-6. doi: 10.1093/nar/gkv400. Epub 2015

Apr 30.

Localize.pytom: a modern webserver for cryo-electron tomography.

Hrabe T(1).

Author information:

(1)Bioinformatics and Systems Biology, Sanford-Burnham Medical Research

Institute, La Jolla, CA 92037, USA thrabe@sanfordburnham.org.

Localize.pytom, available through http://localize.pytom.org is a webserver for

the localize module in the PyTom package. It is a free website and open to all

users and there is no login requirement. The server accepts tomograms as they are

imaged and reconstructed by Cryo-Electron Tomography (CET) and returns densities

and coordinates of candidate-macromolecules in the tomogram. Localization of

macromolecules in cryo-electron tomograms is one of the key procedures to unravel

structural features of imaged macromolecules. Positions of localized molecules

are further used for structural analysis by single particle procedures such as

fine alignment, averaging and classification. Accurate localization can be

furthermore used to generate molecular atlases of whole cells. Localization uses

a cross-correlation-based score and requires a reference volume as input. A

reference can either be a previously detected macromolecular structure or

extrapolated on the server from a specific PDB chain. Users have the option to

use either coarse or fine angular sampling strategies based on uniformly

distributed rotations and to accurately compensate for the CET common 'Missing

Wedge' artefact during sampling. After completion, all candidate macromolecules

cut out from the tomogram are available for download. Their coordinates are

stored and available in XML format, which can be easily integrated into

successive analysis steps in other software. A pre-computed average of the first

one hundred macromolecules is also available for immediate download, and the user

has the option to further analyse the average, based on the detected score

distribution in a novel web-density viewer.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv400

PMCID: PMC4489234

PMID: 25934806 [Indexed for MEDLINE]

520. Nucleic Acids Res. 2015 Jul 1;43(W1):W552-9. doi: 10.1093/nar/gkv399. Epub 2015

May 1.

Pathways with PathWhiz.

Pon A(1), Jewison T(1), Su Y(1), Liang Y(1), Knox C(1), Maciejewski A(1), Wilson

M(1), Wishart DS(2).

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, Alberta, T6G

2E8, Canada. (2)Department of Computing Science, University of Alberta, Edmonton,

Alberta, T6G 2E8, Canada Department of Biological Science, University of Alberta,

Edmonton, Alberta, T6G 2E8, Canada National Institute for Nanotechnology, 11421

Saskatchewan Drive, Edmonton, Alberta, T6G 2M9, Canada david.wishart@ualberta.ca.

PathWhiz (http://smpdb.ca/pathwhiz) is a web server designed to create colourful,

visually pleasing and biologically accurate pathway diagrams that are both

machine-readable and interactive. As a web server, PathWhiz is accessible from

almost any place and compatible with essentially any operating system. It also

houses a public library of pathways and pathway components that can be easily

viewed and expanded upon by its users. PathWhiz allows users to readily generate

biologically complex pathways by using a specially designed drawing palette to

quickly render metabolites (including automated structure generation), proteins

(including quaternary structures, covalent modifications and cofactors), nucleic

acids, membranes, subcellular structures, cells, tissues and organs. Both

small-molecule and protein/gene pathways can be constructed by combining multiple

pathway processes such as reactions, interactions, binding events and transport

activities. PathWhiz's pathway replication and propagation functions allow for

existing pathways to be used to create new pathways or for existing pathways to

be automatically propagated across species. PathWhiz pathways can be saved in

BioPAX, SBGN-ML and SBML data exchange formats, as well as PNG, PWML, HTML image

map or SVG images that can be viewed offline or explored using PathWhiz's

interactive viewer. PathWhiz has been used to generate over 700 pathway diagrams

for a number of popular databases including HMDB, DrugBank and SMPDB.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv399

PMCID: PMC4489271

PMID: 25934797 [Indexed for MEDLINE]

521. Nucleic Acids Res. 2015 Jul 1;43(W1):W314-9. doi: 10.1093/nar/gkv314. Epub 2015

Apr 23.

Stock-based detection of protein oligomeric states in jsPISA.

Krissinel E(1).

Author information:

(1)CCP4, Research Complex at Harwell, Didcot, OX11 0FA, UK

eugene.krissinel@stfc.ac.uk.

A new version of the popular software PISA for the analysis of macromolecular

interfaces and identification of biological assemblies (complexes) from

macromolecular crystal structures is presented. The new web server jsPISA has a

substantially improved user interface, based on modern JavaScript technologies,

and also new elements of analysis: assembly stock and interaction radar. The new

elements help interpretation of PISA results in difficult and ambiguous cases,

for example, when the oligomeric state depends on protein concentration, or when

the biologically relevant interaction is weak and cannot be easily discriminated

from superficial crystal contacts. jsPISA is maintained by CCP4 at

http://www.ccp4.ac.uk/pisa. There are no login requirements for using the server.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv314

PMCID: PMC4489313

PMID: 25908787 [Indexed for MEDLINE]

522. Nucleic Acids Res. 2015 Jul 1;43(W1):W117-21. doi: 10.1093/nar/gkv384. Epub 2015

Apr 20.

Babelomics 5.0: functional interpretation for new generations of genomic data.

Alonso R(1), Salavert F(2), Garcia-Garcia F(3), Carbonell-Caballero J(3), Bleda

M(4), Garcia-Alonso L(3), Sanchis-Juan A(5), Perez-Gil D(5), Marin-Garcia P(5),

Sanchez R(6), Cubuk C(3), Hidalgo MR(3), Amadoz A(3), Hernansaiz-Ballesteros

RD(3), Alemán A(2), Tarraga J(3), Montaner D(3), Medina I(7), Dopazo J(8).

Author information:

(1)Computational Genomics Department, Centro de Investigación Príncipe Felipe

(CIPF), Valencia, 46012, Spain Computational Genomics Chair, Bull-CIPF, Valencia,

46012, Spain. (2)Computational Genomics Department, Centro de Investigación

Príncipe Felipe (CIPF), Valencia, 46012, Spain Bioinformatics of Rare Diseases

(BIER), CIBER de Enfermedades Raras (CIBERER), Valencia, 46012, Spain.

(3)Computational Genomics Department, Centro de Investigación Príncipe Felipe

(CIPF), Valencia, 46012, Spain. (4)Department of Medicine, University of

Cambridge, School of Clinical Medicine, Addenbrooke's Hospital, Hills Road,

Cambridge CB2 0QQ, UK. (5)Fundación Investigación Clínico de Valencia-INCLIVA,

Valencia, 46010, Spain. (6)Computational Genomics Department, Centro de

Investigación Príncipe Felipe (CIPF), Valencia, 46012, Spain Functional Genomics

Node, (INB) at CIPF, Valencia, 46012, Spain. (7)HPC Services, University of

Cambridge, Cambridge, CB3 0RB UK. (8)Computational Genomics Department, Centro de

Investigación Príncipe Felipe (CIPF), Valencia, 46012, Spain Computational

Genomics Chair, Bull-CIPF, Valencia, 46012, Spain Bioinformatics of Rare Diseases

(BIER), CIBER de Enfermedades Raras (CIBERER), Valencia, 46012, Spain Functional

Genomics Node, (INB) at CIPF, Valencia, 46012, Spain jdopazo@cipf.es.

Babelomics has been running for more than one decade offering a user-friendly

interface for the functional analysis of gene expression and genomic data. Here

we present its fifth release, which includes support for Next Generation

Sequencing data including gene expression (RNA-seq), exome or genome

resequencing. Babelomics has simplified its interface, being now more intuitive.

Improved visualization options, such as a genome viewer as well as an interactive

network viewer, have been implemented. New technical enhancements at both, client

and server sides, makes the user experience faster and more dynamic. Babelomics

offers user-friendly access to a full range of methods that cover: (i) primary

data analysis, (ii) a variety of tests for different experimental designs and

(iii) different enrichment and network analysis algorithms for the interpretation

of the results of such tests in the proper functional context. In addition to the

public server, local copies of Babelomics can be downloaded and installed.

Babelomics is freely available at: http://www.babelomics.org.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv384

PMCID: PMC4489263

PMID: 25897133 [Indexed for MEDLINE]

523. Nucleic Acids Res. 2015 Jul 1;43(W1):W251-7. doi: 10.1093/nar/gkv380. Epub 2015

Apr 20.

MetaboAnalyst 3.0--making metabolomics more meaningful.

Xia J(1), Sinelnikov IV(2), Han B(2), Wishart DS(3).

Author information:

(1)Institute of Parasitology, and Department of Animal Science, McGill

University, Ste. Ann de Bellevue, QC H9X 3V9, Canada Department of Microbiology

and Immunology, McGill University, Montreal, QC H3A 2B4, Canada

jeff.xia@mcgill.ca. (2)Department of Computing Science, University of Alberta,

Edmonton, AB T6G 2E8, Canada. (3)Department of Computing Science, University of

Alberta, Edmonton, AB T6G 2E8, Canada Department of Biological Sciences,

University of Alberta, Edmonton, AB T6G 2E9, Canada National Institute for

Nanotechnology, 11421 Saskatchewan Drive, Edmonton, AB T6G 2M9, Canada.

MetaboAnalyst (www.metaboanalyst.ca) is a web server designed to permit

comprehensive metabolomic data analysis, visualization and interpretation. It

supports a wide range of complex statistical calculations and high quality

graphical rendering functions that require significant computational resources.

First introduced in 2009, MetaboAnalyst has experienced more than a 50X growth in

user traffic (>50 000 jobs processed each month). In order to keep up with the

rapidly increasing computational demands and a growing number of requests to

support translational and systems biology applications, we performed a

substantial rewrite and major feature upgrade of the server. The result is

MetaboAnalyst 3.0. By completely re-implementing the MetaboAnalyst suite using

the latest web framework technologies, we have been able substantially improve

its performance, capacity and user interactivity. Three new modules have also

been added including: (i) a module for biomarker analysis based on the

calculation of receiver operating characteristic curves; (ii) a module for sample

size estimation and power analysis for improved planning of metabolomics studies

and (iii) a module to support integrative pathway analysis for both genes and

metabolites. In addition, popular features found in existing modules have been

significantly enhanced by upgrading the graphical output, expanding the compound

libraries and by adding support for more diverse organisms.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv380

PMCID: PMC4489235

PMID: 25897128 [Indexed for MEDLINE]

524. Nucleic Acids Res. 2015 Jul 1;43(W1):W543-6. doi: 10.1093/nar/gkv385. Epub 2015

Apr 20.

The iceLogo web server and SOAP service for determining protein consensus

sequences.

Maddelein D(1), Colaert N(1), Buchanan I(1), Hulstaert N(1), Gevaert K(1),

Martens L(2).

Author information:

(1)Department of Medical Protein Research, VIB, A. Baertsoenkaai 3, B-9000 Ghent,

Belgium Department of Biochemistry, Faculty of Medicine and Health Sciences,

Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium. (2)Department of

Medical Protein Research, VIB, A. Baertsoenkaai 3, B-9000 Ghent, Belgium

Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent

University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium lennart.martens@ugent.be.

The iceLogo web server and SOAP service implement the previously published

iceLogo algorithm. iceLogo builds on probability theory to visualize protein

consensus sequences in a format resembling sequence logos. Peptide sequences are

compared against a reference sequence set that can be tailored to the studied

system and the used protocol. As such, not only over- but also underrepresented

residues can be visualized in a statistically sound manner, which further allows

the user to easily analyse and interpret conserved sequence patterns in proteins.

The web application and SOAP service can be found free and open to all users

without the need for a login on http://iomics.ugent.be/icelogoserver/main.html.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv385

PMCID: PMC4489316

PMID: 25897125 [Indexed for MEDLINE]

525. Nucleic Acids Res. 2015 Jul 1;43(W1):W356-61. doi: 10.1093/nar/gkv368. Epub 2015

Apr 20.

pyDockSAXS: protein-protein complex structure by SAXS and computational docking.

Jiménez-García B(1), Pons C(2), Svergun DI(3), Bernadó P(4), Fernández-Recio

J(5).

Author information:

(1)Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona

Supercomputing Center, 08034 Barcelona, Spain. (2)Department of Computer Science

and Engineering, University of Minnesota, Minneapolis, MN 55455, USA. (3)European

Molecular Biology Laboratory, Hamburg Outstation, 22603 Hamburg, Germany.

(4)Centre de Biochimie Structurale, INSERM U1054, CNRS UMR 5048, Université

Montpellier 1 and 2, F-34090 Montpellier, France. (5)Joint BSC-CRG-IRB Research

Program in Computational Biology, Barcelona Supercomputing Center, 08034

Barcelona, Spain juanf@bsc.es.

Structural characterization of protein-protein interactions at molecular level is

essential to understand biological processes and identify new therapeutic

opportunities. However, atomic resolution structural techniques cannot keep pace

with current advances in interactomics. Low-resolution structural techniques,

such as small-angle X-ray scattering (SAXS), can be applied at larger scale, but

they miss atomic details. For efficient application to protein-protein complexes,

low-resolution information can be combined with theoretical methods that provide

energetic description and atomic details of the interactions. Here we present the

pyDockSAXS web server (http://life.bsc.es/pid/pydocksaxs) that provides an

automatic pipeline for modeling the structure of a protein-protein complex from

SAXS data. The method uses FTDOCK to generate rigid-body docking models that are

subsequently evaluated by a combination of pyDock energy-based scoring function

and their capacity to describe SAXS data. The only required input files are

structural models for the interacting partners and a SAXS curve. The server

automatically provides a series of structural models for the complex, sorted by

the pyDockSAXS scoring function. The user can also upload a previously computed

set of docking poses, which opens the possibility to filter the docking solutions

by potential interface residues or symmetry restraints. The server is freely

available to all users without restriction.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv368

PMCID: PMC4489248

PMID: 25897115 [Indexed for MEDLINE]

526. Nucleic Acids Res. 2015 Jul 1;43(W1):W343-8. doi: 10.1093/nar/gkv357. Epub 2015

Apr 20.

RBO Aleph: leveraging novel information sources for protein structure prediction.

Mabrouk M(1), Putz I(1), Werner T(1), Schneider M(1), Neeb M(1), Bartels P(1),

Brock O(2).

Author information:

(1)Robotics and Biology Laboratory, Department of Electrical Engineering and

Computer Science, Technische Universität Berlin, Marchstraße 23, 10587 Berlin,

Germany. (2)Robotics and Biology Laboratory, Department of Electrical Engineering

and Computer Science, Technische Universität Berlin, Marchstraße 23, 10587

Berlin, Germany oliver.brock@tu-berlin.de.

RBO Aleph is a novel protein structure prediction web server for template-based

modeling, protein contact prediction and ab initio structure prediction. The

server has a strong emphasis on modeling difficult protein targets for which

templates cannot be detected. RBO Aleph's unique features are (i) the use of

combined evolutionary and physicochemical information to perform residue-residue

contact prediction and (ii) leveraging this contact information effectively in

conformational space search. RBO Aleph emerged as one of the leading approaches

to ab initio protein structure prediction and contact prediction during the most

recent Critical Assessment of Protein Structure Prediction experiment (CASP11,

2014). In addition to RBO Aleph's main focus on ab initio modeling, the server

also provides state-of-the-art template-based modeling services. Based on

template availability, RBO Aleph switches automatically between template-based

modeling and ab initio prediction based on the target protein sequence,

facilitating use especially for non-expert users. The RBO Aleph web server offers

a range of tools for visualization and data analysis, such as the visualization

of predicted models, predicted contacts and the estimated prediction error along

the model's backbone. The server is accessible at

http://compbio.robotics.tu-berlin.de/rbo\_aleph/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv357

PMCID: PMC4489312

PMID: 25897112 [Indexed for MEDLINE]

527. Nucleic Acids Res. 2015 Jul 1;43(W1):W571-5. doi: 10.1093/nar/gkv354. Epub 2015

Apr 16.

CellWhere: graphical display of interaction networks organized on subcellular

localizations.

Zhu L(1), Malatras A(2), Thorley M(2), Aghoghogbe I(3), Mer A(4), Duguez S(2),

Butler-Browne G(2), Voit T(2), Duddy W(5).

Author information:

(1)Bioinformatics Department, Bielefeld University, Bielefeld, D-33501, Germany

Center for Research in Myology, Sorbonne Universités, UPMC Univ Paris 06, INSERM

UMRS975, CNRS FRE3617, 47 Boulevard de l'hôpital, 75013 Paris, France. (2)Center

for Research in Myology, Sorbonne Universités, UPMC Univ Paris 06, INSERM

UMRS975, CNRS FRE3617, 47 Boulevard de l'hôpital, 75013 Paris, France. (3)Center

for Research in Myology, Sorbonne Universités, UPMC Univ Paris 06, INSERM

UMRS975, CNRS FRE3617, 47 Boulevard de l'hôpital, 75013 Paris, France

Orthopaedics and Musculoskeletal Science, University College London, London, WC1E

6BT, UK. (4)Medical Epidemiology and Biostatistics, Karolinska Institute,

Stockholm, SE-17177, Sweden. (5)Center for Research in Myology, Sorbonne

Universités, UPMC Univ Paris 06, INSERM UMRS975, CNRS FRE3617, 47 Boulevard de

l'hôpital, 75013 Paris, France william.duddy@upmc.fr.

Given a query list of genes or proteins, CellWhere produces an interactive

graphical display that mimics the structure of a cell, showing the local

interaction network organized into subcellular locations. This user-friendly tool

helps in the formulation of mechanistic hypotheses by enabling the experimental

biologist to explore simultaneously two elements of functional context: (i)

protein subcellular localization and (ii) protein-protein interactions or gene

functional associations. Subcellular localization terms are obtained from public

sources (the Gene Ontology and UniProt-together containing several thousand such

terms) then mapped onto a smaller number of CellWhere localizations. These

localizations include all major cell compartments, but the user may modify the

mapping as desired. Protein-protein interaction listings, and their associated

evidence strength scores, are obtained from the Mentha interactome server, or

power-users may upload a pre-made network produced using some other interactomics

tool. The Cytoscape.js JavaScript library is used in producing the graphical

display. Importantly, for a protein that has been observed at multiple

subcellular locations, users may prioritize the visual display of locations that

are of special relevance to their research domain. CellWhere is at

http://cellwhere-myology.rhcloud.com.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv354

PMCID: PMC4489307

PMID: 25883154 [Indexed for MEDLINE]

528. Nucleic Acids Res. 2015 Jul 1;43(W1):W527-34. doi: 10.1093/nar/gkv344. Epub 2015

Apr 16.

QmRLFS-finder: a model, web server and stand-alone tool for prediction and

analysis of R-loop forming sequences.

Jenjaroenpun P(1), Wongsurawat T(1), Yenamandra SP(1), Kuznetsov VA(2).

Author information:

(1)Department of Genome and Gene Expression Data Analysis, Bioinformatics

Institute, Agency for Science, Technology and Research (A\*STAR), 30 Biopolis

street, #07-01, Singapore 138671. (2)Department of Genome and Gene Expression

Data Analysis, Bioinformatics Institute, Agency for Science, Technology and

Research (A\*STAR), 30 Biopolis street, #07-01, Singapore 138671

vladimirk@bii.a-star.edu.sg.

Erratum in

Nucleic Acids Res. 2015 Nov 16;43(20):10081.

The possible formation of three-stranded RNA and DNA hybrid structures (R-loops)

in thousands of functionally important guanine-rich genic and inter-genic regions

could suggest their involvement in transcriptional regulation and even

development of diseases. Here, we introduce the first freely available R-loop

prediction program called Quantitative Model of R-loop Forming Sequence (RLFS)

finder (QmRLFS-finder), which predicts RLFSs in nucleic acid sequences based on

experimentally supported structural models of RLFSs. QmRLFS-finder operates via a

web server or a stand-alone command line tool. This tool identifies and

visualizes RLFS coordinates from any natural or artificial DNA or RNA input

sequences and creates standards-compliant output files for further annotation and

analysis. QmRLFS-finder demonstrates highly accurate predictions of the detected

RLFSs, proposing new perspective to further discoveries in R-loop biology,

biotechnology and molecular therapy. QmRLFS-finder is freely available at

http://rloop.bii.a-star.edu.sg/?pg=qmrlfs-finder.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv344

PMCID: PMC4489302

PMID: 25883153 [Indexed for MEDLINE]

529. Nucleic Acids Res. 2015 Jul 1;43(W1):W395-400. doi: 10.1093/nar/gkv343. Epub 2015

Apr 16.

ENCoM server: exploring protein conformational space and the effect of mutations

on protein function and stability.

Frappier V(1), Chartier M(1), Najmanovich RJ(2).

Author information:

(1)Department of Biochemistry, Faculty of Medicine and Health Sciences,

University of Sherbrooke, Sherbrooke, Quebec, J1H 5N4, Canada. (2)Department of

Biochemistry, Faculty of Medicine and Health Sciences, University of Sherbrooke,

Sherbrooke, Quebec, J1H 5N4, Canada rafael.najmanovich@usherbrooke.ca.

ENCoM is a coarse-grained normal mode analysis method recently introduced that

unlike previous such methods is unique in that it accounts for the nature of

amino acids. The inclusion of this layer of information was shown to improve

conformational space sampling and apply for the first time a coarse-grained

normal mode analysis method to predict the effect of single point mutations on

protein dynamics and thermostability resulting from vibrational entropy changes.

Here we present a web server that allows non-technical users to have access to

ENCoM calculations to predict the effect of mutations on thermostability and

dynamics as well as to generate geometrically realistic conformational ensembles.

The server is accessible at: http://bcb.med.usherbrooke.ca/encom.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv343

PMCID: PMC4489264

PMID: 25883149 [Indexed for MEDLINE]

530. Nucleic Acids Res. 2015 Jul 1;43(W1):W174-81. doi: 10.1093/nar/gkv342. Epub 2015

Apr 16.

I-TASSER server: new development for protein structure and function predictions.

Yang J(1), Zhang Y(2).

Author information:

(1)Department of Computational Medicine and Bioinformatics, University of

Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA School of

Mathematical Sciences and LPMC, Nankai University, Tianjin, 300071, PR China

yangjy@nankai.edu.cn. (2)Department of Computational Medicine and Bioinformatics,

University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA

Department of Biological Chemistry, University of Michigan, 100 Washtenaw Avenue,

Ann Arbor, MI 48109-2218, USA zhng@umich.edu.

The I-TASSER server (http://zhanglab.ccmb.med.umich.edu/I-TASSER) is an online

resource for automated protein structure prediction and structure-based function

annotation. In I-TASSER, structural templates are first recognized from the PDB

using multiple threading alignment approaches. Full-length structure models are

then constructed by iterative fragment assembly simulations. The functional

insights are finally derived by matching the predicted structure models with

known proteins in the function databases. Although the server has been widely

used for various biological and biomedical investigations, numerous comments and

suggestions have been reported from the user community. In this article, we

summarize recent developments on the I-TASSER server, which were designed to

address the requirements from the user community and to increase the accuracy of

modeling predictions. Focuses have been made on the introduction of new methods

for atomic-level structure refinement, local structure quality estimation and

biological function annotations. We expect that these new developments will

improve the quality of the I-TASSER server and further facilitate its use by the

community for high-resolution structure and function prediction.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv342

PMCID: PMC4489253

PMID: 25883148 [Indexed for MEDLINE]

531. Nucleic Acids Res. 2015 Jul 1;43(W1):W7-14. doi: 10.1093/nar/gkv318. Epub 2015

Apr 16.

GUIDANCE2: accurate detection of unreliable alignment regions accounting for the

uncertainty of multiple parameters.

Sela I(1), Ashkenazy H(1), Katoh K(2), Pupko T(3).

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel-Aviv University, Tel-Aviv 6997801, Israel. (2)Immunology Frontier

Research Center, Osaka University, Suita, Osaka 565-0871, Japan Computational

Biology Research Center, The National Institute of Advanced Industrial Science

and Technology (AIST), Tokyo 135-0064, Japan. (3)Department of Cell Research and

Immunology, George S. Wise Faculty of Life Sciences, Tel-Aviv University,

Tel-Aviv 6997801, Israel talp@post.tau.ac.il.

Inference of multiple sequence alignments (MSAs) is a critical part of

phylogenetic and comparative genomics studies. However, from the same set of

sequences different MSAs are often inferred, depending on the methodologies used

and the assumed parameters. Much effort has recently been devoted to improving

the ability to identify unreliable alignment regions. Detecting such unreliable

regions was previously shown to be important for downstream analyses relying on

MSAs, such as the detection of positive selection. Here we developed GUIDANCE2, a

new integrative methodology that accounts for: (i) uncertainty in the process of

indel formation, (ii) uncertainty in the assumed guide tree and (iii) co-optimal

solutions in the pairwise alignments, used as building blocks in progressive

alignment algorithms. We compared GUIDANCE2 with seven methodologies to detect

unreliable MSA regions using extensive simulations and empirical benchmarks. We

show that GUIDANCE2 outperforms all previously developed methodologies.

Furthermore, GUIDANCE2 also provides a set of alternative MSAs which can be

useful for downstream analyses. The novel algorithm is implemented as a

web-server, available at: http://guidance.tau.ac.il.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv318

PMCID: PMC4489236

PMID: 25883146 [Indexed for MEDLINE]

532. Nucleic Acids Res. 2015 Jul 1;43(W1):W306-13. doi: 10.1093/nar/gkv359. Epub 2015

Apr 16.

AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein

structures.

Zambrano R(1), Jamroz M(2), Szczasiuk A(2), Pujols J(1), Kmiecik S(3), Ventura

S(4).

Author information:

(1)Institut de Biotecnologia i Biomedicina and Departament de Bioquímica i

Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain.

(2)University of Warsaw, Faculty of Chemistry, Pasteura 1, Warsaw, Poland.

(3)University of Warsaw, Faculty of Chemistry, Pasteura 1, Warsaw, Poland

sekmi@chem.uw.edu.pl. (4)Institut de Biotecnologia i Biomedicina and Departament

de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona,

Bellaterra, 08193, Spain salvador.ventura@uab.es.

Protein aggregation underlies an increasing number of disorders and constitutes a

major bottleneck in the development of therapeutic proteins. Our present

understanding on the molecular determinants of protein aggregation has

crystalized in a series of predictive algorithms to identify aggregation-prone

sites. A majority of these methods rely only on sequence. Therefore, they find

difficulties to predict the aggregation properties of folded globular proteins,

where aggregation-prone sites are often not contiguous in sequence or buried

inside the native structure. The AGGRESCAN3D (A3D) server overcomes these

limitations by taking into account the protein structure and the experimental

aggregation propensity scale from the well-established AGGRESCAN method. Using

the A3D server, the identified aggregation-prone residues can be virtually

mutated to design variants with increased solubility, or to test the impact of

pathogenic mutations. Additionally, A3D server enables to take into account the

dynamic fluctuations of protein structure in solution, which may influence

aggregation propensity. This is possible in A3D Dynamic Mode that exploits the

CABS-flex approach for the fast simulations of flexibility of globular proteins.

The A3D server can be accessed at http://biocomp.chem.uw.edu.pl/A3D/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv359

PMCID: PMC4489226

PMID: 25883144 [Indexed for MEDLINE]

533. Nucleic Acids Res. 2015 Jul 1;43(W1):W220-4. doi: 10.1093/nar/gkv335. Epub 2015

Apr 16.

webSDA: a web server to simulate macromolecular diffusional association.

Yu X(1), Martinez M(2), Gable AL(2), Fuller JC(2), Bruce NJ(2), Richter S(2),

Wade RC(3).

Author information:

(1)Molecular and Cellular Modeling group, Heidelberg Institute for Theoretical

Studies (HITS), Heidelberg, Baden-Württemberg, Germany Rebecca.Wade@h-its.org.

(2)Molecular and Cellular Modeling group, Heidelberg Institute for Theoretical

Studies (HITS), Heidelberg, Baden-Württemberg, Germany. (3)Molecular and Cellular

Modeling group, Heidelberg Institute for Theoretical Studies (HITS), Heidelberg,

Baden-Württemberg, Germany Center for Molecular Biology (ZMBH), DKFZ-ZMBH

Alliance, Heidelberg University, Heidelberg, Baden-Württemberg, Germany

Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University,

Heidelberg, Baden-Württemberg, Germany Rebecca.Wade@h-its.org.

Macromolecular interactions play a crucial role in biological systems. Simulation

of diffusional association (SDA) is a software for carrying out Brownian dynamics

simulations that can be used to study the interactions between two or more

biological macromolecules. webSDA allows users to run Brownian dynamics

simulations with SDA to study bimolecular association and encounter complex

formation, to compute association rate constants, and to investigate

macromolecular crowding using atomically detailed macromolecular structures.

webSDA facilitates and automates the use of the SDA software, and offers

user-friendly visualization of results. webSDA currently has three modules: 'SDA

docking' to generate structures of the diffusional encounter complexes of two

macromolecules, 'SDA association' to calculate bimolecular diffusional

association rate constants, and 'SDA multiple molecules' to simulate the

diffusive motion of hundreds of macromolecules. webSDA is freely available to all

users and there is no login requirement. webSDA is available at

http://mcm.h-its.org/webSDA/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv335

PMCID: PMC4489311

PMID: 25883142 [Indexed for MEDLINE]

534. Nucleic Acids Res. 2015 Jul 1;43(W1):W389-94. doi: 10.1093/nar/gkv332. Epub 2015

Apr 16.

JPred4: a protein secondary structure prediction server.

Drozdetskiy A(1), Cole C(1), Procter J(1), Barton GJ(2).

Author information:

(1)Division of Computational Biology, College of Life Sciences, University of

Dundee, Dundee, DD1 5EH, UK. (2)Division of Computational Biology, College of

Life Sciences, University of Dundee, Dundee, DD1 5EH, UK g.j.barton@dundee.ac.uk.

JPred4 (http://www.compbio.dundee.ac.uk/jpred4) is the latest version of the

popular JPred protein secondary structure prediction server which provides

predictions by the JNet algorithm, one of the most accurate methods for secondary

structure prediction. In addition to protein secondary structure, JPred also

makes predictions of solvent accessibility and coiled-coil regions. The JPred

service runs up to 94 000 jobs per month and has carried out over 1.5 million

predictions in total for users in 179 countries. The JPred4 web server has been

re-implemented in the Bootstrap framework and JavaScript to improve its design,

usability and accessibility from mobile devices. JPred4 features higher accuracy,

with a blind three-state (α-helix, β-strand and coil) secondary structure

prediction accuracy of 82.0% while solvent accessibility prediction accuracy has

been raised to 90% for residues <5% accessible. Reporting of results is enhanced

both on the website and through the optional email summaries and batch submission

results. Predictions are now presented in SVG format with options to view full

multiple sequence alignments with and without gaps and insertions. Finally, the

help-pages have been updated and tool-tips added as well as step-by-step

tutorials.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv332

PMCID: PMC4489285

PMID: 25883141 [Indexed for MEDLINE]

535. Nucleic Acids Res. 2015 Jul 1;43(W1):W200-7. doi: 10.1093/nar/gkv353. Epub 2015

Apr 16.

FAF-Drugs3: a web server for compound property calculation and chemical library

design.

Lagorce D(1), Sperandio O(1), Baell JB(2), Miteva MA(1), Villoutreix BO(3).

Author information:

(1)Université Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques In

Silico, Paris 75013, France Inserm U973, Molécules Thérapeutiques In Silico,

Paris 75013, France. (2)Medicinal Chemistry, Monash Institute of Pharmaceutical

Sciences, Monash University, Parkville, Victoria 3052, Australia. (3)Université

Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques In Silico, Paris

75013, France Inserm U973, Molécules Thérapeutiques In Silico, Paris 75013,

France bruno.villoutreix@inserm.fr.

Drug attrition late in preclinical or clinical development is a serious economic

problem in the field of drug discovery. These problems can be linked, in part, to

the quality of the compound collections used during the hit generation stage and

to the selection of compounds undergoing optimization. Here, we present

FAF-Drugs3, a web server that can be used for drug discovery and chemical biology

projects to help in preparing compound libraries and to assist decision-making

during the hit selection/lead optimization phase. Since it was first described in

2006, FAF-Drugs has been significantly modified. The tool now applies an enhanced

structure curation procedure, can filter or analyze molecules with user-defined

or eight predefined physicochemical filters as well as with several simple ADMET

(absorption, distribution, metabolism, excretion and toxicity) rules. In

addition, compounds can be filtered using an updated list of 154 hand-curated

structural alerts while Pan Assay Interference compounds (PAINS) and other,

generally unwanted groups are also investigated. FAF-Drugs3 offers access to

user-friendly html result pages and the possibility to download all computed

data. The server requires as input an SDF file of the compounds; it is open to

all users and can be accessed without registration at

http://fafdrugs3.mti.univ-paris-diderot.fr.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv353

PMCID: PMC4489254

PMID: 25883137 [Indexed for MEDLINE]

536. Nucleic Acids Res. 2015 Jul 1;43(W1):W244-50. doi: 10.1093/nar/gkv347. Epub 2015

Apr 15.

PheNetic: network-based interpretation of molecular profiling data.

De Maeyer D(1), Weytjens B(1), Renkens J(2), De Raedt L(2), Marchal K(3).

Author information:

(1)Dept. of Microbial and Molecular Systems, KULeuven, Leuven, 3000, Belgium

Dept. of Information Technology (INTEC, iMINDS), U.Ghent, Ghent, 9052, Belgium.

(2)Dept. of Computer Science, KULeuven, Leuven, 3000, Belgium. (3)Dept. of

Microbial and Molecular Systems, KULeuven, Leuven, 3000, Belgium Dept. of

Information Technology (INTEC, iMINDS), U.Ghent, Ghent, 9052, Belgium Dept. of

Plant Biotechnology and Bioinformatics, U.Ghent, Ghent, 9052, Belgium

kathleen.marchal@intec.ugent.be.

Molecular profiling experiments have become standard in current wet-lab

practices. Classically, enrichment analysis has been used to identify biological

functions related to these experimental results. Combining molecular profiling

results with the wealth of currently available interactomics data, however,

offers the opportunity to identify the molecular mechanism behind an observed

molecular phenotype. In this paper, we therefore introduce 'PheNetic', a

user-friendly web server for inferring a sub-network based on probabilistic

logical querying. PheNetic extracts from an interactome, the sub-network that

best explains genes prioritized through a molecular profiling experiment.

Depending on its run mode, PheNetic searches either for a regulatory mechanism

that gave explains to the observed molecular phenotype or for the pathways

(in)activated in the molecular phenotype. The web server provides access to a

large number of interactomes, making sub-network inference readily applicable to

a wide variety of organisms. The inferred sub-networks can be interactively

visualized in the browser. PheNetic's method and use are illustrated using an

example analysis of differential expression results of ampicillin treated

Escherichia coli cells. The PheNetic web service is available at

http://bioinformatics.intec.ugent.be/phenetic/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv347

PMCID: PMC4489255

PMID: 25878035 [Indexed for MEDLINE]

537. Nucleic Acids Res. 2015 Jul 1;43(W1):W301-5. doi: 10.1093/nar/gkv346. Epub 2015

Apr 15.

NGS-eval: NGS Error analysis and novel sequence VAriant detection tooL.

May A(1), Abeln S(2), Buijs MJ(3), Heringa J(2), Crielaard W(3), Brandt BW(4).

Author information:

(1)Department of Preventive Dentistry, Academic Centre for Dentistry Amsterdam

(ACTA), University of Amsterdam and VU University Amsterdam, Amsterdam, The

Netherlands Centre for Integrative Bioinformatics (IBIVU), VU University

Amsterdam, Amsterdam, The Netherlands. (2)Centre for Integrative Bioinformatics

(IBIVU), VU University Amsterdam, Amsterdam, The Netherlands AIMMS Amsterdam

Institute for Molecules Medicines and Systems, VU University Amsterdam,

Amsterdam, The Netherlands. (3)Department of Preventive Dentistry, Academic

Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and VU University

Amsterdam, Amsterdam, The Netherlands. (4)Department of Preventive Dentistry,

Academic Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and VU

University Amsterdam, Amsterdam, The Netherlands b.brandt@acta.nl.

Massively parallel sequencing of microbial genetic markers (MGMs) is used to

uncover the species composition in a multitude of ecological niches. These

sequencing runs often contain a sample with known composition that can be used to

evaluate the sequencing quality or to detect novel sequence variants. With

NGS-eval, the reads from such (mock) samples can be used to (i) explore the

differences between the reads and their references and to (ii) estimate the

sequencing error rate. This tool maps these reads to references and calculates as

well as visualizes the different types of sequencing errors. Clearly, sequencing

errors can only be accurately calculated if the reference sequences are correct.

However, even with known strains, it is not straightforward to select the correct

references from databases. We previously analysed a pyrosequencing dataset from a

mock sample to estimate sequencing error rates and detected sequence variants in

our mock community, allowing us to obtain an accurate error estimation. Here, we

demonstrate the variant detection and error analysis capability of NGS-eval with

Illumina MiSeq reads from the same mock community. While tailored towards the

field of metagenomics, this server can be used for any type of MGM-based reads.

NGS-eval is available at http://www.ibi.vu.nl/programs/ngsevalwww/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv346

PMCID: PMC4489229

PMID: 25878034 [Indexed for MEDLINE]

538. Nucleic Acids Res. 2015 Jul 1;43(W1):W225-30. doi: 10.1093/nar/gkv309. Epub 2015

Apr 8.

WAXSiS: a web server for the calculation of SAXS/WAXS curves based on

explicit-solvent molecular dynamics.

Knight CJ(1), Hub JS(2).

Author information:

(1)Institute for Microbiology and Genetics, Georg-August-University Göttingen,

Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany. (2)Institute for Microbiology

and Genetics, Georg-August-University Göttingen, Justus-von-Liebig-Weg 11, 37077

Göttingen, Germany jhub@gwdg.de.

Small- and wide-angle X-ray scattering (SWAXS) has evolved into a powerful tool

to study biological macromolecules in solution. The interpretation of SWAXS

curves requires their accurate predictions from structural models. Such

predictions are complicated by scattering contributions from the hydration layer

and by effects from thermal fluctuations. Here, we describe the new web server

WAXSiS (WAXS in solvent) that computes SWAXS curves based on explicit-solvent

all-atom molecular dynamics (MD) simulations (http://waxsis.uni-goettingen.de/).

The MD simulations provide a realistic model for both the hydration layer and the

excluded solvent, thereby avoiding any solvent-related fitting parameters, while

naturally accounting for thermal fluctuations.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv309

PMCID: PMC4489308

PMID: 25855813 [Indexed for MEDLINE]

539. Nucleic Acids Res. 2015 Jul 1;43(W1):W448-54. doi: 10.1093/nar/gkv306. Epub 2015

Apr 8.

MTiOpenScreen: a web server for structure-based virtual screening.

Labbé CM(1), Rey J(2), Lagorce D(1), Vavruša M(2), Becot J(1), Sperandio O(1),

Villoutreix BO(1), Tufféry P(2), Miteva MA(3).

Author information:

(1)Université Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques In

Silico, INSERM UMR-S 973, Paris, France INSERM, U973, Paris, France.

(2)Université Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques In

Silico, INSERM UMR-S 973, Paris, France INSERM, U973, Paris, France RPBS, 75205

Paris, France. (3)Université Paris Diderot, Sorbonne Paris Cité, Molécules

Thérapeutiques In Silico, INSERM UMR-S 973, Paris, France INSERM, U973, Paris,

France maria.miteva@univ-paris-diderot.fr.

Open screening endeavors play and will play a key role to facilitate the

identification of new bioactive compounds in order to foster innovation and to

improve the effectiveness of chemical biology and drug discovery processes. In

this line, we developed the new web server MTiOpenScreen dedicated to small

molecule docking and virtual screening. It includes two services, MTiAutoDock and

MTiOpenScreen, allowing performing docking into a user-defined binding site or

blind docking using AutoDock 4.2 and automated virtual screening with AutoDock

Vina. MTiOpenScreen provides valuable starting collections for screening, two

in-house prepared drug-like chemical libraries containing 150 000 PubChem

compounds: the Diverse-lib containing diverse molecules and the iPPI-lib enriched

in molecules likely to inhibit protein-protein interactions. In addition,

MTiOpenScreen offers users the possibility to screen up to 5000 small molecules

selected outside our two libraries. The predicted binding poses and energies of

up to 1000 top ranked ligands can be downloaded. In this way, MTiOpenScreen

enables researchers to apply virtual screening using different chemical libraries

on traditional or more challenging protein targets such as protein-protein

interactions. The MTiOpenScreen web server is free and open to all users at

http://bioserv.rpbs.univ-paris-diderot.fr/services/MTiOpenScreen/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv306

PMCID: PMC4489289

PMID: 25855812 [Indexed for MEDLINE]

540. Nucleic Acids Res. 2015 Jul 1;43(W1):W24-9. doi: 10.1093/nar/gkv317. Epub 2015

Apr 8.

SANSparallel: interactive homology search against Uniprot.

Somervuo P(1), Holm L(2).

Author information:

(1)Institute of Biotechnology, University of Helsinki, PO Box 65, Finland

Department of Biosciences, University of Helsinki, PO Box 65, Finland.

(2)Institute of Biotechnology, University of Helsinki, PO Box 65, Finland

Department of Biosciences, University of Helsinki, PO Box 65, Finland

liisa.holm@helsinki.fi.

Proteins evolve by mutations and natural selection. The network of sequence

similarities is a rich source for mining homologous relationships that inform on

protein structure and function. There are many servers available to browse the

network of homology relationships but one has to wait up to a minute for results.

The SANSparallel webserver provides protein sequence database searches with

immediate response and professional alignment visualization by third-party

software. The output is a list, pairwise alignment or stacked alignment of

sequence-similar proteins from Uniprot, UniRef90/50, Swissprot or Protein Data

Bank. The stacked alignments are viewed in Jalview or as sequence logos. The

database search uses the suffix array neighborhood search (SANS) method, which

has been re-implemented as a client-server, improved and parallelized. The method

is extremely fast and as sensitive as BLAST above 50% sequence identity.

Benchmarks show that the method is highly competitive compared to previously

published fast database search programs: UBLAST, DIAMOND, LAST, LAMBDA,

RAPSEARCH2 and BLAT. The web server can be accessed interactively or

programmatically at http://ekhidna2.biocenter.helsinki.fi/cgi-bin/sans/sans.cgi.

It can be used to make protein functional annotation pipelines more efficient,

and it is useful in interactive exploration of the detailed evidence supporting

the annotation of particular proteins of interest.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv317

PMCID: PMC4489265

PMID: 25855811 [Indexed for MEDLINE]

541. Nucleic Acids Res. 2015 Jul 1;43(W1):W3-6. doi: 10.1093/nar/gkv310. Epub 2015 Apr

8.

TCS: a web server for multiple sequence alignment evaluation and phylogenetic

reconstruction.

Chang JM(1), Di Tommaso P(2), Lefort V(3), Gascuel O(3), Notredame C(4).

Author information:

(1)Comparative Bioinformatics, Bioinformatics and Genomics Program, Centre for

Genomic Regulation (CRG), Dr Aiguader 88, 08003 Barcelona, Spain Universitat

Pompeu Fabra (UPF), 08003 Barcelona, Spain Institute of Human Genetics (IGH), UPR

1142, CNRS, 34396 Montpellier, France. (2)Comparative Bioinformatics,

Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Dr

Aiguader 88, 08003 Barcelona, Spain Universitat Pompeu Fabra (UPF), 08003

Barcelona, Spain. (3)Institut de Biologie Computationnelle, LIRMM, UMR 5506: CNRS

& Université de Montpellier, France. (4)Comparative Bioinformatics,

Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Dr

Aiguader 88, 08003 Barcelona, Spain Universitat Pompeu Fabra (UPF), 08003

Barcelona, Spain cedric.notredame@crg.eu.

This article introduces the Transitive Consistency Score (TCS) web server; a

service making it possible to estimate the local reliability of protein multiple

sequence alignments (MSAs) using the TCS index. The evaluation can be used to

identify the aligned positions most likely to contain structurally analogous

residues and also most likely to support an accurate phylogenetic reconstruction.

The TCS scoring scheme has been shown to be accurate predictor of structural

alignment correctness among commonly used methods. It has also been shown to

outperform common filtering schemes like Gblocks or trimAl when doing MSA

post-processing prior to phylogenetic tree reconstruction. The web server is

available from http://tcoffee.crg.cat/tcs.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv310

PMCID: PMC4489230

PMID: 25855806 [Indexed for MEDLINE]

542. Nucleic Acids Res. 2015 Jul 1;43(W1):W169-73. doi: 10.1093/nar/gkv236. Epub 2015

Mar 27.

IntFOLD: an integrated server for modelling protein structures and functions from

amino acid sequences.

McGuffin LJ(1), Atkins JD(2), Salehe BR(2), Shuid AN(2), Roche DB(3).

Author information:

(1)School of Biological Sciences, University of Reading, Reading, RG6 6AS, UK

l.j.mcguffin@reading.ac.uk. (2)School of Biological Sciences, University of

Reading, Reading, RG6 6AS, UK. (3)Institut de Biologie Computationnelle, LIRMM,

CNRS, Université de Montpellier, Montpellier 34095, France Centre de Recherches

de Biochimie Macromoléculaire, CNRS- UMR 5237, Montpellier 34293, France.

IntFOLD is an independent web server that integrates our leading methods for

structure and function prediction. The server provides a simple unified interface

that aims to make complex protein modelling data more accessible to life

scientists. The server web interface is designed to be intuitive and integrates a

complex set of quantitative data, so that 3D modelling results can be viewed on a

single page and interpreted by non-expert modellers at a glance. The only

required input to the server is an amino acid sequence for the target protein.

Here we describe major performance and user interface updates to the server,

which comprises an integrated pipeline of methods for: tertiary structure

prediction, global and local 3D model quality assessment, disorder prediction,

structural domain prediction, function prediction and modelling of protein-ligand

interactions. The server has been independently validated during numerous CASP

(Critical Assessment of Techniques for Protein Structure Prediction) experiments,

as well as being continuously evaluated by the CAMEO (Continuous Automated Model

Evaluation) project. The IntFOLD server is available at:

http://www.reading.ac.uk/bioinf/IntFOLD/.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv236

PMCID: PMC4489238

PMID: 25820431 [Indexed for MEDLINE]

543. Nucleic Acids Res. 2015 Jul 1;43(W1):W122-7. doi: 10.1093/nar/gkv253. Epub 2015

Mar 26.

RiceNet v2: an improved network prioritization server for rice genes.

Lee T(1), Oh T(2), Yang S(1), Shin J(1), Hwang S(1), Kim CY(1), Kim H(1), Shim

H(1), Shim JE(1), Ronald PC(3), Lee I(4).

Author information:

(1)Department of Biotechnology, College of Life Sciences and Biotechnology,

Yonsei University, Seoul, Korea. (2)The Joint Bioenergy Institute, Emeryville CA

and Department of Plant Pathology and the Genome Center, University of

California, Davis, CA 95616, USA. (3)The Joint Bioenergy Institute, Emeryville CA

and Department of Plant Pathology and the Genome Center, University of

California, Davis, CA 95616, USA insuklee@yonsei.ac.kr. (4)Department of

Biotechnology, College of Life Sciences and Biotechnology, Yonsei University,

Seoul, Korea insuklee@yonsei.ac.kr.

Rice is the most important staple food crop and a model grass for studies of

bioenergy crops. We previously published a genome-scale functional network server

called RiceNet, constructed by integrating diverse genomics data and demonstrated

the use of the network in genetic dissection of rice biotic stress responses and

its usefulness for other grass species. Since the initial construction of the

network, there has been a significant increase in the amount of publicly

available rice genomics data. Here, we present an updated network prioritization

server for Oryza sativa ssp. japonica, RiceNet v2

(http://www.inetbio.org/ricenet), which provides a network of 25 765 genes (70.1%

of the coding genome) and 1 775 000 co-functional links. Ricenet v2 also provides

two complementary methods for network prioritization based on: (i) network direct

neighborhood and (ii) context-associated hubs. RiceNet v2 can use genes of the

related subspecies O. sativa ssp. indica and the reference plant Arabidopsis for

versatility in generating hypotheses. We demonstrate that RiceNet v2 effectively

identifies candidate genes involved in rice root/shoot development and defense

responses, demonstrating its usefulness for the grass research community.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gkv253

PMCID: PMC4489288

PMID: 25813048 [Indexed for MEDLINE]

544. PLoS One. 2015 Jun 26;10(6):e0130834. doi: 10.1371/journal.pone.0130834.

eCollection 2015.

"Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and

Management of Web Servers Supporting Online Studies.

Lange K(1), Kühn S(2), Filevich E(2).

Author information:

(1)Independent Researcher, Berlin, Germany. (2)Department of Lifespan Psychology,

Max Planck Institute for Human Development, Berlin, Germany.

Erratum in

PLoS One. 2015;10(7):e0134073.

We present here "Just Another Tool for Online Studies" (JATOS): an open source,

cross-platform web application with a graphical user interface (GUI) that greatly

simplifies setting up and communicating with a web server to host online studies

that are written in JavaScript. JATOS is easy to install in all three major

platforms (Microsoft Windows, Mac OS X, and Linux), and seamlessly pairs with a

database for secure data storage. It can be installed on a server or locally,

allowing researchers to try the application and feasibility of their studies

within a browser environment, before engaging in setting up a server. All

communication with the JATOS server takes place via a GUI (with no need to use a

command line interface), making JATOS an especially accessible tool for

researchers without a strong IT background. We describe JATOS' main features and

implementation and provide a detailed tutorial along with example studies to help

interested researchers to set up their online studies. JATOS can be found under

the Internet address: www.jatos.org.

DOI: 10.1371/journal.pone.0130834

PMCID: PMC4482716

PMID: 26114751 [Indexed for MEDLINE]

545. J Chem Inf Model. 2015 Jun 22;55(6):1261-70. doi: 10.1021/ci500577m. Epub 2015

May 18.

A Novel Cylindrical Representation for Characterizing Intrinsic Properties of

Protein Sequences.

Yu JF(1,)(2), Dou XH(1), Wang HB(1), Sun X(2), Zhao HY(3), Wang JH(1,)(4).

Author information:

(1)†Shandong Provincial Key Laboratory of Functional Macromolecular Biophysics,

Institute of Biophysics, Dezhou University, Dezhou 253023, China. (2)‡State Key

Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China.

(3)§Department of Genetics and Computational Biology, QIMR Berghofer Medical

Research Institute, Brisbane, Queensland 4000, Australia. (4)∥College of Physics

and Electronic Information, Dezhou University, Dezhou 253023, China.

The composition and sequence order of amino acid residues are the two most

important characteristics to describe a protein sequence. Graphical

representations facilitate visualization of biological sequences and produce

biologically useful numerical descriptors. In this paper, we propose a novel

cylindrical representation by placing the 20 amino acid residue types in a circle

and sequence positions along the z axis. This representation allows visualization

of the composition and sequence order of amino acids at the same time. Ten

numerical descriptors and one weighted numerical descriptor have been developed

to quantitatively describe intrinsic properties of protein sequences on the basis

of the cylindrical model. Their applications to similarity/dissimilarity analysis

of nine ND5 proteins indicated that these numerical descriptors are more

effective than several classical numerical matrices. Thus, the cylindrical

representation obtained here provides a new useful tool for visualizing and

charactering protein sequences. An online server is available at

http://biophy.dzu.edu.cn:8080/CNumD/input.jsp .

DOI: 10.1021/ci500577m

PMID: 25945398 [Indexed for MEDLINE]

546. Sci Rep. 2015 Jun 18;5:10184. doi: 10.1038/srep10184.

iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide

position-specific propensity.

Xu Y(1), Ding YX(1), Ding J(1), Lei YH(1), Wu LY(2), Deng NY(3).

Author information:

(1)Department of Information and Computer Science, University of Science and

Technology Beijing, Beijing 100083, China. (2)Institute of Applied Mathematics,

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing

100190, China. (3)College of Science, China Agricultural University, Beijing

100083, China.

Lysine succinylation in protein is one type of post-translational modifications

(PTMs). Succinylation is associated with some diseases and succinylated sites

data just has been found in recent years in experiments. It is highly desired to

develop computational methods to identify the candidate proteins and their sites.

In view of this, a new predictor called iSuc-PseAAC was proposed by incorporating

the peptide position-specific propensity into the general form of pseudo amino

acid composition. The accuracy is 79.94%, sensitivity 51.07%, specificity 89.42%

and MCC 0.431 in leave-one-out cross validation with support vector machine

algorithm. It demonstrated by rigorous leave-one-out on stringent benchmark

dataset that the new predictor is quite promising and may become a useful high

throughput tool in this area. Meanwhile a user-friendly web-server for

iSuc-PseAAC is accessible at http://app.aporc.org/iSuc-PseAAC/. Users can easily

obtain their desired results without the need to understand the complicated

mathematical equations presented in this paper just for its integrity.

DOI: 10.1038/srep10184

PMCID: PMC4471726

PMID: 26084794 [Indexed for MEDLINE]

547. Int J Mol Sci. 2015 Jun 16;16(6):13829-49. doi: 10.3390/ijms160613829.

Identifying Similar Patterns of Structural Flexibility in Proteins by Disorder

Prediction and Dynamic Programming.

Petrovich A(1), Borne A(2), Uversky VN(3,)(4,)(5,)(6), Xue B(7).

Author information:

(1)Department of Physics, University of South Florida, Tampa, FL 33620, USA.

aidan1@mail.usf.edu. (2)Department of Cell Biology, Microbiology, and Molecular

Biology, School of Natural Sciences and Mathematics, College of Arts and

Sciences, University of South Florida, Tampa, FL 33620, USA. bornea@mail.usf.edu.

(3)Department of Molecular Medicine and USF Health Byrd Alzheimer's Research

Institute, Morsani College of Medicine, University of South Florida, Tampa, FL

33620, USA. vuversky@health.usf.edu. (4)Institute for Biological Instrumentation,

Russian Academy of Sciences, Pushchino, Moscow Region 142290, Russian.

vuversky@health.usf.edu. (5)Department of Biology, Faculty of Science, King

Abdulaziz University, P.O. Box 80203, Jeddah 21589, Saudi Arabia.

vuversky@health.usf.edu. (6)Laboratory of Structural Dynamics, Stability and

Folding of Proteins, Institute of Cytology, Russian Academy of Sciences, St.

Petersburg 194064, Russian. vuversky@health.usf.edu. (7)Department of Cell

Biology, Microbiology, and Molecular Biology, School of Natural Sciences and

Mathematics, College of Arts and Sciences, University of South Florida, Tampa, FL

33620, USA. binxue@usf.edu.

Computational methods are prevailing in identifying protein intrinsic disorder.

The results from predictors are often given as per-residue disorder scores. The

scores describe the disorder propensity of amino acids of a protein and can be

further represented as a disorder curve. Many proteins share similar patterns in

their disorder curves. The similar patterns are often associated with similar

functions and evolutionary origins. Therefore, finding and characterizing

specific patterns of disorder curves provides a unique and attractive perspective

of studying the function of intrinsically disordered proteins. In this study, we

developed a new computational tool named IDalign using dynamic programming. This

tool is able to identify similar patterns among disorder curves, as well as to

present the distribution of intrinsic disorder in query proteins. The

disorder-based information generated by IDalign is significantly different from

the information retrieved from classical sequence alignments. This tool can also

be used to infer functions of disordered regions and disordered proteins. The web

server of IDalign is available at (http://labs.cas.usf.edu/bioinfo/service.html).

DOI: 10.3390/ijms160613829

PMCID: PMC4490526

PMID: 26086829 [Indexed for MEDLINE]

548. PLoS One. 2015 Jun 16;10(6):e0129635. doi: 10.1371/journal.pone.0129635.

eCollection 2015.

Computational Identification of Protein Pupylation Sites by Using Profile-Based

Composition of k-Spaced Amino Acid Pairs.

Hasan MM(1), Zhou Y(1), Lu X(1), Li J(2), Song J(3), Zhang Z(1).

Author information:

(1)State Key Laboratory of Agrobiotechnology, College of Biological Sciences,

China Agricultural University, Beijing, 100193, China. (2)Advanced Analytics

Institute and Centre for Health Technologies, University of Technology, Sydney,

81 Broadway, NSW 2007, Australia. (3)National Engineering Laboratory for

Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin

Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin,

300308, China; Monash Bioinformatics Platform and Department of Biochemistry and

Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800,

Australia.

Prokaryotic proteins are regulated by pupylation, a type of post-translational

modification that contributes to cellular function in bacterial organisms. In

pupylation process, the prokaryotic ubiquitin-like protein (Pup) tagging is

functionally analogous to ubiquitination in order to tag target proteins for

proteasomal degradation. To date, several experimental methods have been

developed to identify pupylated proteins and their pupylation sites, but these

experimental methods are generally laborious and costly. Therefore, computational

methods that can accurately predict potential pupylation sites based on protein

sequence information are highly desirable. In this paper, a novel predictor

termed as pbPUP has been developed for accurate prediction of pupylation sites.

In particular, a sophisticated sequence encoding scheme [i.e. the profile-based

composition of k-spaced amino acid pairs (pbCKSAAP)] is used to represent the

sequence patterns and evolutionary information of the sequence fragments

surrounding pupylation sites. Then, a Support Vector Machine (SVM) classifier is

trained using the pbCKSAAP encoding scheme. The final pbPUP predictor achieves an

AUC value of 0.849 in 10-fold cross-validation tests and outperforms other

existing predictors on a comprehensive independent test dataset. The proposed

method is anticipated to be a helpful computational resource for the prediction

of pupylation sites. The web server and curated datasets in this study are freely

available at http://protein.cau.edu.cn/pbPUP/.

DOI: 10.1371/journal.pone.0129635

PMCID: PMC4469302

PMID: 26080082 [Indexed for MEDLINE]

549. Bioinformatics. 2015 Jun 15;31(12):i133-41. doi: 10.1093/bioinformatics/btv242.

Finding optimal interaction interface alignments between biological complexes.

Cui X(1), Naveed H(1), Gao X(1).

Author information:

(1)Computer, Electrical and Mathematical Sciences and Engineering Division, King

Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi

Arabia.

MOTIVATION: Biological molecules perform their functions through interactions

with other molecules. Structure alignment of interaction interfaces between

biological complexes is an indispensable step in detecting their structural

similarities, which are key S: to understanding their evolutionary histories and

functions. Although various structure alignment methods have been developed to

successfully access the similarities of protein structures or certain types of

interaction interfaces, existing alignment tools cannot directly align arbitrary

types of interfaces formed by protein, DNA or RNA molecules. Specifically, they

require a ': blackbox preprocessing ': to standardize interface types and chain

identifiers. Yet their performance is limited and sometimes unsatisfactory.

RESULTS: Here we introduce a novel method, PROSTA-inter, that automatically

determines and aligns interaction interfaces between two arbitrary types of

complex structures. Our method uses sequentially remote fragments to search for

the optimal superimposition. The optimal residue matching problem is then

formulated as a maximum weighted bipartite matching problem to detect the optimal

sequence order-independent alignment. Benchmark evaluation on all non-redundant

protein -: DNA complexes in PDB shows significant performance improvement of our

method over TM-align and iAlign (with the ': blackbox preprocessing ': ). Two

case studies where our method discovers, for the first time, structural

similarities between two pairs of functionally related protein -: DNA complexes

are presented. We further demonstrate the power of our method on detecting

structural similarities between a protein -: protein complex and a protein -: RNA

complex, which is biologically known as a protein -: RNA mimicry case.

AVAILABILITY AND IMPLEMENTATION: The PROSTA-inter web-server is publicly

available at http://www.cbrc.kaust.edu.sa/prosta/.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv242

PMCID: PMC4765866

PMID: 26072475 [Indexed for MEDLINE]

550. Bioinformatics. 2015 Jun 15;31(12):i116-23. doi: 10.1093/bioinformatics/btv235.

Large-scale model quality assessment for improving protein tertiary structure

prediction.

Cao R(1), Bhattacharya D(1), Adhikari B(1), Li J(1), Cheng J(2).

Author information:

(1)Computer Science Department, University of Missouri, Columbia, Missouri,

65211, USA, Informatics Institute, University of Missouri, Columbia, Missouri,

65211, USA and C. Bond Life Science Center, University of Missouri, Columbia,

Missouri, 65211, USA. (2)Computer Science Department, University of Missouri,

Columbia, Missouri, 65211, USA, Informatics Institute, University of Missouri,

Columbia, Missouri, 65211, USA and C. Bond Life Science Center, University of

Missouri, Columbia, Missouri, 65211, USA Computer Science Department, University

of Missouri, Columbia, Missouri, 65211, USA, Informatics Institute, University of

Missouri, Columbia, Missouri, 65211, USA and C. Bond Life Science Center,

University of Missouri, Columbia, Missouri, 65211, USA Computer Science

Department, University of Missouri, Columbia, Missouri, 65211, USA, Informatics

Institute, University of Missouri, Columbia, Missouri, 65211, USA and C. Bond

Life Science Center, University of Missouri, Columbia, Missouri, 65211, USA.

MOTIVATION: Sampling structural models and ranking them are the two major

challenges of protein structure prediction. Traditional protein structure

prediction methods generally use one or a few quality assessment (QA) methods to

select the best-predicted models, which cannot consistently select relatively

better models and rank a large number of models well.

RESULTS: Here, we develop a novel large-scale model QA method in conjunction with

model clustering to rank and select protein structural models. It unprecedentedly

applied 14 model QA methods to generate consensus model rankings, followed by

model refinement based on model combination (i.e. averaging). Our experiment

demonstrates that the large-scale model QA approach is more consistent and robust

in selecting models of better quality than any individual QA method. Our method

was blindly tested during the 11th Critical Assessment of Techniques for Protein

Structure Prediction (CASP11) as MULTICOM group. It was officially ranked third

out of all 143 human and server predictors according to the total scores of the

first models predicted for 78 CASP11 protein domains and second according to the

total scores of the best of the five models predicted for these domains.

MULTICOM's outstanding performance in the extremely competitive 2014 CASP11

experiment proves that our large-scale QA approach together with model clustering

is a promising solution to one of the two major problems in protein structure

modeling.

AVAILABILITY AND IMPLEMENTATION: The web server is available at:

http://sysbio.rnet.missouri.edu/multicom\_cluster/human/.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv235

PMCID: PMC4553833

PMID: 26072473 [Indexed for MEDLINE]

551. Bioinformatics. 2015 Jun 15;31(12):2046-8. doi: 10.1093/bioinformatics/btv087.

Epub 2015 Feb 11.

SOMP: web server for in silico prediction of sites of metabolism for drug-like

compounds.

Rudik A(1), Dmitriev A(1), Lagunin A(2), Filimonov D(1), Poroikov V(2).

Author information:

(1)Laboratory for Structure-Function Based Drug Design, Institute of Biomedical

Chemistry, 10/8 Pogodinskaya Str., Moscow, 119121, Russia and Medicobiological

Faculty, Pirogov Russian National Research Medical University, 1 Ostrovityanova

Str., Moscow, 117997, Russia. (2)Laboratory for Structure-Function Based Drug

Design, Institute of Biomedical Chemistry, 10/8 Pogodinskaya Str., Moscow,

119121, Russia and Medicobiological Faculty, Pirogov Russian National Research

Medical University, 1 Ostrovityanova Str., Moscow, 117997, Russia Laboratory for

Structure-Function Based Drug Design, Institute of Biomedical Chemistry, 10/8

Pogodinskaya Str., Moscow, 119121, Russia and Medicobiological Faculty, Pirogov

Russian National Research Medical University, 1 Ostrovityanova Str., Moscow,

117997, Russia.

A new freely available web server site of metabolism predictor to predict the

sites of metabolism (SOM) based on the structural formula of chemicals has been

developed. It is based on the analyses of 'structure-SOM' relationships using a

Bayesian approach and labelled multilevel neighbourhoods of atoms descriptors to

represent the structures of over 1000 metabolized xenobiotics. The server allows

predicting SOMs that are catalysed by 1A2, 2C9, 2C19, 2D6 and 3A4 isoforms of

cytochrome P450 and enzymes of the UDP-glucuronosyltransferase family. The

average invariant accuracy of prediction that was calculated for the training

sets (using leave-one-out cross-validation) and evaluation sets is 0.9 and 0.95,

respectively.AVAILABILITY AND IMPLEMENTATION: Freely available on the web at

http://www.way2drug.com/SOMP.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv087

PMID: 25777527 [Indexed for MEDLINE]

552. Bioinformatics. 2015 Jun 15;31(12):1966-73. doi: 10.1093/bioinformatics/btv100.

Epub 2015 Feb 19.

Extending P450 site-of-metabolism models with region-resolution data.

Zaretzki JM(1), Browning MR(1), Hughes TB(1), Swamidass SJ(1).

Author information:

(1)Department of Pathology and Immunology, Washington University School of

Medicine, St. Louis, MO 63130, USA.

MOTIVATION: Cytochrome P450s are a family of enzymes responsible for the

metabolism of approximately 90% of FDA-approved drugs. Medicinal chemists often

want to know which atoms of a molecule-its metabolized sites-are oxidized by

Cytochrome P450s in order to modify their metabolism. Consequently, there are

several methods that use literature-derived, atom-resolution data to train models

that can predict a molecule's sites of metabolism. There is, however, much more

data available at a lower resolution, where the exact site of metabolism is not

known, but the region of the molecule that is oxidized is known. Until now, no

site-of-metabolism models made use of region-resolution data.

RESULTS: Here, we describe XenoSite-Region, the first reported method for

training site-of-metabolism models with region-resolution data. Our approach uses

the Expectation Maximization algorithm to train a site-of-metabolism model.

Region-resolution metabolism data was simulated from a large site-of-metabolism

dataset, containing 2000 molecules with 3400 metabolized and 30 000

un-metabolized sites and covering nine Cytochrome P450 isozymes. When training on

the same molecules (but with only region-level information), we find that this

approach yields models almost as accurate as models trained with atom-resolution

data. Moreover, we find that atom-resolution trained models are more accurate

when also trained with region-resolution data from additional molecules. Our

approach, therefore, opens up a way to extend the applicable domain of

site-of-metabolism models into larger regions of chemical space. This meets a

critical need in drug development by tapping into underutilized data commonly

available in most large drug companies.

AVAILABILITY AND IMPLEMENTATION: The algorithm, data and a web server are

available at http://swami.wustl.edu/xregion.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv100

PMID: 25697821 [Indexed for MEDLINE]

553. J Cheminform. 2015 Jun 15;7:25. doi: 10.1186/s13321-015-0077-3. eCollection 2015.

ChemDIS: a chemical-disease inference system based on chemical-protein

interactions.

Tung CW(1).

Author information:

(1)School of Pharmacy, College of Pharmacy, Kaohsiung Medical University, 100

Shih-Chuan 1st Road, Kaohsiung, 80708 Taiwan ; Ph.D. Program in Toxicology,

College of Pharmacy, Kaohsiung Medical University, Kaohsiung, 80708 Taiwan ;

National Environmental Health Research Center, National Health Research

Institutes, Miaoli County, 35053 Taiwan.

BACKGROUND: The characterization of toxicities associated with environmental and

industrial chemicals is required for risk assessment. However, we lack the

toxicological data for a large portion of chemicals due to the high cost of

experiments for a huge number of chemicals. The development of computational

methods for identifying potential risks associated with chemicals is desirable

for generating testable hypothesis to accelerate the hazard identification

process.

RESULTS: A chemical-disease inference system named ChemDIS was developed to

facilitate hazard identification for chemicals. The chemical-protein interactions

from a large database STITCH and protein-disease relationship from disease

ontology and disease ontology lite were utilized for chemical-protein-disease

inferences. Tools with user-friendly interfaces for enrichment analysis of

functions, pathways and diseases were implemented and integrated into ChemDIS. An

analysis on maleic acid and sibutramine showed that ChemDIS could be a useful

tool for the identification of potential functions, pathways and diseases

affected by poorly characterized chemicals.

CONCLUSIONS: ChemDIS is an integrated chemical-disease inference system for

poorly characterized chemicals with potentially affected functions and pathways

for experimental validation. ChemDIS server is freely accessible at

http://cwtung.kmu.edu.tw/chemdis.

DOI: 10.1186/s13321-015-0077-3

PMCID: PMC4466364

PMID: 26078786

554. J Theor Biol. 2015 Jun 7;374:60-5. doi: 10.1016/j.jtbi.2015.03.029. Epub 2015 Apr

2.

Accurate in silico identification of protein succinylation sites using an

iterative semi-supervised learning technique.

Zhao X(1), Ning Q(2), Chai H(2), Ma Z(3).

Author information:

(1)School of Computer Science and Information Technology, Northeast Normal

University, Changchun, 130117, China. Electronic address: zhaoxw303@nenu.edu.cn.

(2)School of Computer Science and Information Technology, Northeast Normal

University, Changchun, 130117, China. (3)Key Laboratory of Intelligent

Information Processing of Jilin Universities, Northeast Normal University,

Changchun 130117, China. Electronic address: zhiqiang.ma967@gmail.com.

As a widespread type of protein post-translational modifications (PTMs),

succinylation plays an important role in regulating protein conformation,

function and physicochemical properties. Compared with the labor-intensive and

time-consuming experimental approaches, computational predictions of

succinylation sites are much desirable due to their convenient and fast speed.

Currently, numerous computational models have been developed to identify PTMs

sites through various types of two-class machine learning algorithms. These

methods require both positive and negative samples for training. However,

designation of the negative samples of PTMs was difficult and if it is not

properly done can affect the performance of computational models dramatically. So

that in this work, we implemented the first application of positive samples only

learning (PSoL) algorithm to succinylation sites prediction problem, which was a

special class of semi-supervised machine learning that used positive samples and

unlabeled samples to train the model. Meanwhile, we proposed a novel

succinylation sites computational predictor called SucPred (succinylation site

predictor) by using multiple feature encoding schemes. Promising results were

obtained by the SucPred predictor with an accuracy of 88.65% using 5-fold cross

validation on the training dataset and an accuracy of 84.40% on the independent

testing dataset, which demonstrated that the positive samples only learning

algorithm presented here was particularly useful for identification of protein

succinylation sites. Besides, the positive samples only learning algorithm can be

applied to build predictors for other types of PTMs sites with ease. A web server

for predicting succinylation sites was developed and was freely accessible at

http://59.73.198.144:8088/SucPred/.

Copyright © 2015 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2015.03.029

PMID: 25843215 [Indexed for MEDLINE]

555. Bioinformatics. 2015 Jun 1;31(11):1854-6. doi: 10.1093/bioinformatics/btv068.

Epub 2015 Feb 1.

Expitope: a web server for epitope expression.

Haase K(1), Raffegerst S(2), Schendel DJ(2), Frishman D(3).

Author information:

(1)Department of Genome Oriented Bioinformatics, Wissenschaftszentrum

Weihenstephan, Technische Universität München, 85354 Freising, Germany, Immune

Monitoring Group, Helmholtz Zentrum Munich, Institute of Molecular Immunology,

81377 München, Germany, Medigene Immunotherapies GmbH a Subsidiary of Medigene

AG, 82152 Planegg, Germany, Helmholtz Zentrum Munich, Institute of Bioinformatics

and Systems Biology, 85764 Neuherberg, Germany and St Petersburg State

Polytechnical University, St Petersburg, 195251, Russia. (2)Department of Genome

Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische

Universität München, 85354 Freising, Germany, Immune Monitoring Group, Helmholtz

Zentrum Munich, Institute of Molecular Immunology, 81377 München, Germany,

Medigene Immunotherapies GmbH a Subsidiary of Medigene AG, 82152 Planegg,

Germany, Helmholtz Zentrum Munich, Institute of Bioinformatics and Systems

Biology, 85764 Neuherberg, Germany and St Petersburg State Polytechnical

University, St Petersburg, 195251, Russia Department of Genome Oriented

Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität

München, 85354 Freising, Germany, Immune Monitoring Group, Helmholtz Zentrum

Munich, Institute of Molecular Immunology, 81377 München, Germany, Medigene

Immunotherapies GmbH a Subsidiary of Medigene AG, 82152 Planegg, Germany,

Helmholtz Zentrum Munich, Institute of Bioinformatics and Systems Biology, 85764

Neuherberg, Germany and St Petersburg State Polytechnical University, St

Petersburg, 195251, Russia. (3)Department of Genome Oriented Bioinformatics,

Wissenschaftszentrum Weihenstephan, Technische Universität München, 85354

Freising, Germany, Immune Monitoring Group, Helmholtz Zentrum Munich, Institute

of Molecular Immunology, 81377 München, Germany, Medigene Immunotherapies GmbH a

Subsidiary of Medigene AG, 82152 Planegg, Germany, Helmholtz Zentrum Munich,

Institute of Bioinformatics and Systems Biology, 85764 Neuherberg, Germany and St

Petersburg State Polytechnical University, St Petersburg, 195251, Russia

Department of Genome Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan,

Technische Universität München, 85354 Freising, Germany, Immune Monitoring Group,

Helmholtz Zentrum Munich, Institute of Molecular Immunology, 81377 München,

Germany, Medigene Immunotherapies GmbH a Subsidiary of Medigene AG, 82152

Planegg, Germany, Helmholtz Zentrum Munich, Institute of Bioinformatics and

Systems Biology, 85764 Neuherberg, Germany and St Petersburg State Polytechnical

University, St Petersburg, 195251, Russia Department of Genome Oriented

Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität

München, 85354 Freising, Germany, Immune Monitoring Group, Helmholtz Zentrum

Munich, Institute of Molecular Immunology, 81377 München, Germany, Medigene

Immunotherapies GmbH a Subsidiary of Medigene AG, 82152 Planegg, Germany,

Helmholtz Zentrum Munich, Institute of Bioinformatics and Systems Biology, 85764

Neuherberg, Germany and St Petersburg State Polytechnical University, St

Petersburg, 195251, Russia.

MOTIVATION: Adoptive T cell therapies based on introduction of new T cell

receptors (TCRs) into patient recipient T cells is a promising new treatment for

various kinds of cancers. A major challenge, however, is the choice of target

antigens. If an engineered TCR can cross-react with self-antigens in healthy

tissue, the side-effects can be devastating. We present the first web server for

assessing epitope sharing when designing new potential lead targets. We enable

the users to find all known proteins containing their peptide of interest. The

web server returns not only exact matches, but also approximate ones, allowing a

number of mismatches of the users choice. For the identified candidate proteins

the expression values in various healthy tissues, representing all vital human

organs, are extracted from RNA Sequencing (RNA-Seq) data as well as from some

cancer tissues as control. All results are returned to the user sorted by a

score, which is calculated using well-established methods and tools for

immunological predictions. It depends on the probability that the epitope is

created by proteasomal cleavage and its affinities to the transporter associated

with antigen processing and the major histocompatibility complex class I alleles.

With this framework, we hope to provide a helpful tool to exclude potential

cross-reactivity in the early stage of TCR selection for use in design of

adoptive T cell immunotherapy.

AVAILABILITY AND IMPLEMENTATION: The Expitope web server can be accessed via

http://webclu.bio.wzw.tum.de/expitope.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please email: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv068

PMID: 25644270 [Indexed for MEDLINE]

556. Bioinformatics. 2015 Jun 1;31(11):1860-2. doi: 10.1093/bioinformatics/btv058.

Epub 2015 Jan 30.

NetExplore: a web server for modeling small network motifs.

Papatsenko D(1), Lemischka IR(2).

Author information:

(1)Department of Regenerative and Developmental Biology, Black Family Stem Cell

Institute andDepartment of Pharmacology and System Therapeutics, Icahn School of

Medicine at Mount Sinai, Systems Biology Center New York, One Gustave L. Levy

Place, New York, NY 10029, USA Department of Regenerative and Developmental

Biology, Black Family Stem Cell Institute andDepartment of Pharmacology and

System Therapeutics, Icahn School of Medicine at Mount Sinai, Systems Biology

Center New York, One Gustave L. Levy Place, New York, NY 10029, USA.

(2)Department of Regenerative and Developmental Biology, Black Family Stem Cell

Institute andDepartment of Pharmacology and System Therapeutics, Icahn School of

Medicine at Mount Sinai, Systems Biology Center New York, One Gustave L. Levy

Place, New York, NY 10029, USA Department of Regenerative and Developmental

Biology, Black Family Stem Cell Institute andDepartment of Pharmacology and

System Therapeutics, Icahn School of Medicine at Mount Sinai, Systems Biology

Center New York, One Gustave L. Levy Place, New York, NY 10029, USA Department of

Regenerative and Developmental Biology, Black Family Stem Cell Institute

andDepartment of Pharmacology and System Therapeutics, Icahn School of Medicine

at Mount Sinai, Systems Biology Center New York, One Gustave L. Levy Place, New

York, NY 10029, USA.

MOTIVATION: Quantitative and qualitative assessment of biological data often

produces small essential recurrent networks, containing 3-5 components called

network motifs. In this context, model solutions for small network motifs

represent very high interest.

RESULTS: Software package NetExplore has been created in order to generate,

classify and analyze solutions for network motifs including up to six network

components. NetExplore allows plotting and visualization of the solution's phase

spaces and bifurcation diagrams.

AVAILABILITY AND IMPLEMENTATION: The current version of NetExplore has been

implemented in Perl-CGI and is accessible at the following locations:

http://line.bioinfolab.net/nex/NetExplore.htm and

http://nex.autosome.ru/nex/NetExplore.htm.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv058

PMCID: PMC4462664

PMID: 25637559 [Indexed for MEDLINE]

557. Bioinformatics. 2015 Jun 1;31(11):1857-9. doi: 10.1093/bioinformatics/btv042.

Epub 2015 Jan 24.

protr/ProtrWeb: R package and web server for generating various numerical

representation schemes of protein sequences.

Xiao N(1), Cao DS(1), Zhu MF(1), Xu QS(1).

Author information:

(1)School of Mathematics and Statistics and School of Pharmaceutical Sciences,

Central South University, Changsha 410083, People's Republic of China.

Amino acid sequence-derived structural and physiochemical descriptors are

extensively utilized for the research of structural, functional, expression and

interaction profiles of proteins and peptides. We developed protr, a

comprehensive R package for generating various numerical representation schemes

of proteins and peptides from amino acid sequence. The package calculates eight

descriptor groups composed of 22 types of commonly used descriptors that include

about 22 700 descriptor values. It allows users to select amino acid properties

from the AAindex database, and use self-defined properties to construct

customized descriptors. For proteochemometric modeling, it calculates six types

of scales-based descriptors derived by various dimensionality reduction methods.

The protr package also integrates the functionality of similarity score

computation derived by protein sequence alignment and Gene Ontology semantic

similarity measures within a list of proteins, and calculates profile-based

protein features based on position-specific scoring matrix. We also developed

ProtrWeb, a user-friendly web server for calculating descriptors presented in the

protr package.AVAILABILITY AND IMPLEMENTATION: The protr package is freely

available from CRAN: http://cran.r-project.org/package=protr, ProtrWeb, is freely

available at http://protrweb.scbdd.com/.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv042

PMID: 25619996 [Indexed for MEDLINE]

558. Bioinformatics. 2015 Jun 1;31(11):1845-7. doi: 10.1093/bioinformatics/btv035.

Epub 2015 Jan 22.

3USS: a web server for detecting alternative 3'UTRs from RNA-seq experiments.

Le Pera L(1), Mazzapioda M(2), Tramontano A(3).

Author information:

(1)Center for Life Nano Science@Sapienza, Istituto Italiano di Tecnologia, Rome,

Italy, Department of Physics, Sapienza University, Rome, Italy and Istituto

Pasteur - Fondazione Cenci Bolognetti, Sapienza University, Rome, Italy Center

for Life Nano Science@Sapienza, Istituto Italiano di Tecnologia, Rome, Italy,

Department of Physics, Sapienza University, Rome, Italy and Istituto Pasteur -

Fondazione Cenci Bolognetti, Sapienza University, Rome, Italy. (2)Center for Life

Nano Science@Sapienza, Istituto Italiano di Tecnologia, Rome, Italy, Department

of Physics, Sapienza University, Rome, Italy and Istituto Pasteur - Fondazione

Cenci Bolognetti, Sapienza University, Rome, Italy. (3)Center for Life Nano

Science@Sapienza, Istituto Italiano di Tecnologia, Rome, Italy, Department of

Physics, Sapienza University, Rome, Italy and Istituto Pasteur - Fondazione Cenci

Bolognetti, Sapienza University, Rome, Italy Center for Life Nano

Science@Sapienza, Istituto Italiano di Tecnologia, Rome, Italy, Department of

Physics, Sapienza University, Rome, Italy and Istituto Pasteur - Fondazione Cenci

Bolognetti, Sapienza University, Rome, Italy Center for Life Nano

Science@Sapienza, Istituto Italiano di Tecnologia, Rome, Italy, Department of

Physics, Sapienza University, Rome, Italy and Istituto Pasteur - Fondazione Cenci

Bolognetti, Sapienza University, Rome, Italy.

Protein-coding genes with multiple alternative polyadenylation sites can generate

mRNA 3'UTR sequences of different lengths, thereby causing the loss or gain of

regulatory elements, which can affect stability, localization and translation

efficiency. 3USS is a web-server developed with the aim of giving

experimentalists the possibility to automatically identify alternative 3 ': UTRs

(shorter or longer with respect to a reference transcriptome), an option that is

not available in standard RNA-seq data analysis procedures. The tool reports as

putative novel the 3 ': UTRs not annotated in available databases. Furthermore,

if data from two related samples are uploaded, common and specific alternative 3

': UTRs are identified and reported by the server.AVAILABILITY AND

IMPLEMENTATION: 3USS is freely available at

http://www.biocomputing.it/3uss\_server.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btv035

PMCID: PMC4443675

PMID: 25617413 [Indexed for MEDLINE]

559. Bioinformatics. 2015 Jun 1;31(11):1869-71. doi: 10.1093/bioinformatics/btv043.

Epub 2015 Jan 22.

CypRules: a rule-based P450 inhibition prediction server.

Shao CY(1), Su BH(1), Tu YS(1), Lin C(1), Lin OA(1), Tseng YJ(2).

Author information:

(1)Graduate Institute of Biomedical Electronics and Bioinformatics and Department

of Computer Science and Information Engineering, National Taiwan University, No.1

Sec.4, Roosevelt Road, Taipei, Taiwan 106. (2)Graduate Institute of Biomedical

Electronics and Bioinformatics and Department of Computer Science and Information

Engineering, National Taiwan University, No.1 Sec.4, Roosevelt Road, Taipei,

Taiwan 106 Graduate Institute of Biomedical Electronics and Bioinformatics and

Department of Computer Science and Information Engineering, National Taiwan

University, No.1 Sec.4, Roosevelt Road, Taipei, Taiwan 106.

Cytochrome P450 (CYPs) are the major enzymes involved in drug metabolism and

bioactivation. Inhibition models were constructed for five of the most popular

enzymes from the CYP superfamily in human liver. The five enzymes chosen for this

study, namely CYP1A2, CYP2D6, CYP2C19, CYP2C9 and CYP3A4, account for 90% of the

xenobiotic and drug metabolism in human body. CYP enzymes can be inhibited or

induced by various drugs or chemical compounds. In this work, a rule-based CYP

inhibition prediction online server, CypRules, was created based on predictive

models generated by the rule-based C5.0 algorithm. CypRules can predict and

provide structural rulesets for CYP inhibition for each compound uploaded to the

server. Capable of fast execution performance, it can be used for virtual

high-throughput screening (VHTS) of a large set of testing compounds.AVAILABILITY

AND IMPLEMENTATION: CypRules is freely accessible at http://cyprules.cmdm.tw/ and

models, descriptor and program files for all compounds are publically available

at http://cyprules.cmdm.tw/sources/sources.rar.

© The Author 2015. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btv043

PMID: 25617412 [Indexed for MEDLINE]

560. Methods. 2015 Jun;79-80:32-40. doi: 10.1016/j.ymeth.2014.10.003. Epub 2014 Oct

13.

Current trend of annotating single nucleotide variation in humans--A case study

on SNVrap.

Li MJ(1), Wang J(2).

Author information:

(1)Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong

Kong, Hong Kong Special Administrative Region, China; Department of Biochemistry,

LKS Faculty of Medicine, The University of Hong Kong, Hong Kong Special

Administrative Region, China; Shenzhen Institute of Research and Innovation, The

University of Hong Kong, Shenzhen, China. (2)Centre for Genomic Sciences, LKS

Faculty of Medicine, The University of Hong Kong, Hong Kong Special

Administrative Region, China; Department of Biochemistry, LKS Faculty of

Medicine, The University of Hong Kong, Hong Kong Special Administrative Region,

China; Shenzhen Institute of Research and Innovation, The University of Hong

Kong, Shenzhen, China. Electronic address: junwen@hku.hk.

As high throughput methods, such as whole genome genotyping arrays, whole exome

sequencing (WES) and whole genome sequencing (WGS), have detected huge amounts of

genetic variants associated with human diseases, function annotation of these

variants is an indispensable step in understanding disease etiology. Large-scale

functional genomics projects, such as The ENCODE Project and Roadmap Epigenomics

Project, provide genome-wide profiling of functional elements across different

human cell types and tissues. With the urgent demands for identification of

disease-causal variants, comprehensive and easy-to-use annotation tool is highly

in demand. Here we review and discuss current progress and trend of the variant

annotation field. Furthermore, we introduce a comprehensive web portal for

annotating human genetic variants. We use gene-based features and the latest

functional genomics datasets to annotate single nucleotide variation (SNVs) in

human, at whole genome scale. We further apply several function prediction

algorithms to annotate SNVs that might affect different biological processes,

including transcriptional gene regulation, alternative splicing,

post-transcriptional regulation, translation and post-translational

modifications. The SNVrap web portal is freely available at

http://jjwanglab.org/snvrap.

Copyright © 2014 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ymeth.2014.10.003

PMID: 25308971 [Indexed for MEDLINE]

561. Mol Inform. 2015 Jun;34(6-7):467-76. doi: 10.1002/minf.201400150. Epub 2015 May

27.

Peptide Binding Prediction to Five Most Frequent HLA-DQ Proteins - a

Proteochemometric Approach.

Dimitrov I(1), Doytchinova I(2).

Author information:

(1)Faculty of Pharmacy, Medical University of Sofia, 2 Dunav str, 1000 Sofia,

Bulgaria tel: +359 2 9236506. (2)Faculty of Pharmacy, Medical University of

Sofia, 2 Dunav str, 1000 Sofia, Bulgaria tel: +359 2 9236506.

idoytchinova@pharmfac.net.

Major histocompatibility complex (MHC) proteins class II, are glycoproteins

binding within the cell to short peptides with foreign origin, called epitopes,

and present them at the cell surface for inspection by T-cells. Apart from

presenting foreign antigens, they are able to present also common self-antigens

and trigger autoimmune diseases as coeliac disease and diabetes mellitus type 1.

The MHC proteins are extremely polymorphic. The polymorphism is located mainly in

the peptide binding site. In the present study, we apply a proteochemometric

approach to derive a model for prediction of peptide binding to human MHC class

II proteins from locus HLA-DQ. Proteochemometrics was applied on 2624 peptides

binding to five most frequent HLA-DQ proteins. The sequences of peptides and

proteins were described by three z-descriptors relating to hydrophobicity, steric

effects and polarity of amino acids. Cross-terms accounting for the

protein-peptide interactions also were included. The derived model was validated

by external test set of 660 peptides and showed rpred (2) =0.808, AUC=0.965,

92.5 % accuracy at threshold of pIC50 =5.3 and average sensitivity of 83 % among

the top 10 % best predicted nonamers. The model is implemented in the server for

MHC binding prediction EpiTOP and is freely available at

http://www.ddg-pharmfac.net/epitop.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

DOI: 10.1002/minf.201400150

PMID: 27490390

562. Nat Protoc. 2015 Jun;10(6):845-58. doi: 10.1038/nprot.2015.053. Epub 2015 May 7.

The Phyre2 web portal for protein modeling, prediction and analysis.

Kelley LA(1), Mezulis S(1), Yates CM(1), Wass MN(1), Sternberg MJ(1).

Author information:

(1)Structural Bioinformatics Group, Imperial College London, London, UK.

Phyre2 is a suite of tools available on the web to predict and analyze protein

structure, function and mutations. The focus of Phyre2 is to provide biologists

with a simple and intuitive interface to state-of-the-art protein bioinformatics

tools. Phyre2 replaces Phyre, the original version of the server for which we

previously published a paper in Nature Protocols. In this updated protocol, we

describe Phyre2, which uses advanced remote homology detection methods to build

3D models, predict ligand binding sites and analyze the effect of amino acid

variants (e.g., nonsynonymous SNPs (nsSNPs)) for a user's protein sequence. Users

are guided through results by a simple interface at a level of detail they

determine. This protocol will guide users from submitting a protein sequence to

interpreting the secondary and tertiary structure of their models, their domain

composition and model quality. A range of additional available tools is described

to find a protein structure in a genome, to submit large number of sequences at

once and to automatically run weekly searches for proteins that are difficult to

model. The server is available at http://www.sbg.bio.ic.ac.uk/phyre2. A typical

structure prediction will be returned between 30 min and 2 h after submission.

DOI: 10.1038/nprot.2015.053

PMCID: PMC5298202

PMID: 25950237 [Indexed for MEDLINE]

Conflict of interest statement: MJES is a Director and shareholder in Equinox

Pharma Ltd which uses bioinformatics and chemoinformatics in drug discovery

research and services.

563. Radiat Oncol J. 2015 Jun;33(2):142-8. doi: 10.3857/roj.2015.33.2.142. Epub 2015

Jun 30.

Development of new on-line statistical program for the Korean Society for

Radiation Oncology.

Song SY(1), Ahn SD(1), Chung WK(2), Shin KH(3), Choi EK(1), Cho KH(4).

Author information:

(1)Department of Radiation Oncology, Asan Medical Center, University of Ulsan

College of Medicine, Seoul, Korea. (2)Department of Radiation Oncology, Kyung Hee

University Hospital at Kangdong, Kyung Hee University School of Medicine, Seoul,

Korea. (3)Department of Radiation Oncology, Seoul National University Hospital,

Seoul National University College of Medicine, Seoul, Korea. (4)Department of

Radiation Oncology, Proton Therapy Center, National Cancer Center Hospital,

Goyang, Korea.

PURPOSE: To develop new on-line statistical program for the Korean Society for

Radiation Oncology (KOSRO) to collect and extract medical data in radiation

oncology more efficiently.

MATERIALS AND METHODS: The statistical program is a web-based program. The

directory was placed in a sub-folder of the homepage of KOSRO and its web address

is http://www.kosro.or.kr/asda. The operating systems server is Linux and the

webserver is the Apache HTTP server. For database (DB) server, MySQL is adopted

and dedicated scripting language is the PHP. Each ID and password are controlled

independently and all screen pages for data input or analysis are made to be

friendly to users. Scroll-down menu is actively used for the convenience of user

and the consistence of data analysis.

RESULTS: Year of data is one of top categories and main topics include human

resource, equipment, clinical statistics, specialized treatment and research

achievement. Each topic or category has several subcategorized topics. Real-time

on-line report of analysis is produced immediately after entering each data and

the administrator is able to monitor status of data input of each hospital.

Backup of data as spread sheets can be accessed by the administrator and be used

for academic works by any members of the KOSRO.

CONCLUSION: The new on-line statistical program was developed to collect data

from nationwide departments of radiation oncology. Intuitive screen and

consistent input structure are expected to promote entering data of member

hospitals and annual statistics should be a cornerstone of advance in radiation

oncology.

DOI: 10.3857/roj.2015.33.2.142

PMCID: PMC4493426

PMID: 26157684

564. Sci Rep. 2015 May 27;5:10576. doi: 10.1038/srep10576.

A statistical framework to predict functional non-coding regions in the human

genome through integrated analysis of annotation data.

Lu Q(1), Hu Y(1), Sun J(1), Cheng Y(2), Cheung KH(3), Zhao H(4).

Author information:

(1)Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA.

(2)Program of Computational Biology and Bioinformatics, Yale University, New

Haven, CT, USA. (3)1] Program of Computational Biology and Bioinformatics, Yale

University, New Haven, CT, USA [2] Yale Center for Medical Informatics, Yale

School of Medicine, New Haven, CT, USA [3] Department of Emergency Medicine, Yale

School of Medicine, New Haven, CT, USA [4] VA Connecticut Healthcare System, West

Haven, CT, USA. (4)1] Department of Biostatistics, Yale School of Public Health,

New Haven, CT, USA [2] Program of Computational Biology and Bioinformatics, Yale

University, New Haven, CT, USA [3] VA Connecticut Healthcare System, West Haven,

CT, USA.

Identifying functional regions in the human genome is a major goal in human

genetics. Great efforts have been made to functionally annotate the human genome

either through computational predictions, such as genomic conservation, or

high-throughput experiments, such as the ENCODE project. These efforts have

resulted in a rich collection of functional annotation data of diverse types that

need to be jointly analyzed for integrated interpretation and annotation. Here we

present GenoCanyon, a whole-genome annotation method that performs unsupervised

statistical learning using 22 computational and experimental annotations thereby

inferring the functional potential of each position in the human genome. With

GenoCanyon, we are able to predict many of the known functional regions. The

ability of predicting functional regions as well as its generalizable statistical

framework makes GenoCanyon a unique and powerful tool for whole-genome

annotation. The GenoCanyon web server is available at

http://genocanyon.med.yale.edu.

DOI: 10.1038/srep10576

PMCID: PMC4444969

PMID: 26015273 [Indexed for MEDLINE]

565. PLoS One. 2015 May 26;10(5):e0128326. doi: 10.1371/journal.pone.0128326.

eCollection 2015.

SEGEL: A Web Server for Visualization of Smoking Effects on Human Lung Gene

Expression.

Xu Y(1), Hu B(1), Alnajm SS(2), Lu Y(1), Huang Y(3), Allen-Gipson D(4), Cheng

F(5).

Author information:

(1)Department of Pharmaceutical Science, College of Pharmacy, University of South

Florida, Tampa, Florida, United States of America. (2)Department of Biology,

Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne,

Germany. (3)Department of Department of Epidemiology and Biostatistics, College

of Public Health, University of South Florida, Tampa, Florida, United States of

America. (4)Department of Pharmaceutical Science, College of Pharmacy, University

of South Florida, Tampa, Florida, United States of America; Division of Allergy

and Clinical Immunology, College of Medicine, University of South Florida, Tampa,

Florida, United States of America. (5)Department of Pharmaceutical Science,

College of Pharmacy, University of South Florida, Tampa, Florida, United States

of America; Department of Department of Epidemiology and Biostatistics, College

of Public Health, University of South Florida, Tampa, Florida, United States of

America.

Cigarette smoking is a major cause of death worldwide resulting in over six

million deaths per year. Cigarette smoke contains complex mixtures of chemicals

that are harmful to nearly all organs of the human body, especially the lungs.

Cigarette smoking is considered the major risk factor for many lung diseases,

particularly chronic obstructive pulmonary diseases (COPD) and lung cancer.

However, the underlying molecular mechanisms of smoking-induced lung injury

associated with these lung diseases still remain largely unknown. Expression

microarray techniques have been widely applied to detect the effects of smoking

on gene expression in different human cells in the lungs. These projects have

provided a lot of useful information for researchers to understand the potential

molecular mechanism(s) of smoke-induced pathogenesis. However, a user-friendly

web server that would allow scientists to fast query these data sets and compare

the smoking effects on gene expression across different cells had not yet been

established. For that reason, we have integrated eight public expression

microarray data sets from trachea epithelial cells, large airway epithelial

cells, small airway epithelial cells, and alveolar macrophage into an online web

server called SEGEL (Smoking Effects on Gene Expression of Lung). Users can query

gene expression patterns across these cells from smokers and nonsmokers by gene

symbols, and find the effects of smoking on the gene expression of lungs from

this web server. Sex difference in response to smoking is also shown. The

relationship between the gene expression and cigarette smoking consumption were

calculated and are shown in the server. The current version of SEGEL web server

contains 42,400 annotated gene probe sets represented on the Affymetrix Human

Genome U133 Plus 2.0 platform. SEGEL will be an invaluable resource for

researchers interested in the effects of smoking on gene expression in the lungs.

The server also provides useful information for drug development against

smoking-related diseases. The SEGEL web server is available online at

http://www.chengfeng.info/smoking\_database.html.

DOI: 10.1371/journal.pone.0128326

PMCID: PMC4444269

PMID: 26010234 [Indexed for MEDLINE]

566. PLoS One. 2015 May 21;10(5):e0127877. doi: 10.1371/journal.pone.0127877.

eCollection 2015.

CleavPredict: A Platform for Reasoning about Matrix Metalloproteinases

Proteolytic Events.

Kumar S(1), Ratnikov BI(1), Kazanov MD(2), Smith JW(1), Cieplak P(1).

Author information:

(1)Sanford Burnham Medical Research Institute, La Jolla, California, United

States of America. (2)Institute for Information Transmission Problems, Russian

Academy of Science, Moscow, Russia.

Erratum in

PLoS One. 2015;10(6):e0131952.

CleavPredict (http://cleavpredict.sanfordburnham.org) is a Web server for

substrate cleavage prediction for matrix metalloproteinases (MMPs). It is

intended as a computational platform aiding the scientific community in reasoning

about proteolytic events. CleavPredict offers in silico prediction of cleavage

sites specific for 11 human MMPs. The prediction method employs the MMP specific

position weight matrices (PWMs) derived from statistical analysis of

high-throughput phage display experimental results. To augment the substrate

cleavage prediction process, CleavPredict provides information about the

structural features of potential cleavage sites that influence proteolysis. These

include: secondary structure, disordered regions, transmembrane domains, and

solvent accessibility. The server also provides information about subcellular

location, co-localization, and co-expression of proteinase and potential

substrates, along with experimentally determined positions of single nucleotide

polymorphism (SNP), and posttranslational modification (PTM) sites in substrates.

All this information will provide the user with perspectives in reasoning about

proteolytic events. CleavPredict is freely accessible, and there is no login

required.

DOI: 10.1371/journal.pone.0127877

PMCID: PMC4440711

PMID: 25996941 [Indexed for MEDLINE]

567. Neurosci Lett. 2015 May 19;595:60-2. doi: 10.1016/j.neulet.2015.03.071. Epub 2015

Apr 7.

The Budapest Reference Connectome Server v2.0.

Szalkai B(1), Kerepesi C(2), Varga B(3), Grolmusz V(4).

Author information:

(1)PIT Bioinformatics Group, Eötvös University, H-1117 Budapest, Hungary.

Electronic address: szalkai@pitgroup.org. (2)PIT Bioinformatics Group, Eötvös

University, H-1117 Budapest, Hungary. Electronic address: kerepesi@pitgroup.org.

(3)PIT Bioinformatics Group, Eötvös University, H-1117 Budapest, Hungary.

Electronic address: balorkany@pitgroup.org. (4)PIT Bioinformatics Group, Eötvös

University, H-1117 Budapest, Hungary; Uratim Ltd., H-1118 Budapest, Hungary.

Electronic address: grolmusz@pitgroup.org.

The connectomes of different human brains are pairwise distinct: we cannot talk

about an abstract "graph of the brain". Two typical connectomes, however, have

quite a few common graph edges that may describe the same connections between the

same cortical areas. The Budapest Reference Connectome Server v2.0 generates the

common edges of the connectomes of 96 distinct cortexes, each with 1015 vertices,

computed from 96 MRI data sets of the Human Connectome Project. The user may set

numerous parameters for the identification and filtering of common edges, and the

graphs are downloadable in both csv and GraphML formats; both formats carry the

anatomical annotations of the vertices, generated by the FreeSurfer program. The

resulting consensus graph is also automatically visualized in a 3D rotating brain

model on the website. The consensus graphs, generated with various parameter

settings, can be used as reference connectomes based on different, independent

MRI images, therefore they may serve as reduced-error, low-noise, robust graph

representations of the human brain. The webserver is available at

http://connectome.pitgroup.org.

Copyright © 2015 Elsevier Ireland Ltd. All rights reserved.

DOI: 10.1016/j.neulet.2015.03.071

PMID: 25862487 [Indexed for MEDLINE]

568. PLoS One. 2015 May 19;10(5):e0127431. doi: 10.1371/journal.pone.0127431.

eCollection 2015.

Quantifying the displacement of mismatches in multiple sequence alignment

benchmarks.

Bawono P(1), van der Velde A(1), Abeln S(1), Heringa J(2).

Author information:

(1)Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam,

Amsterdam, The Netherlands. (2)Centre for Integrative Bioinformatics (IBIVU), VU

University Amsterdam, Amsterdam, The Netherlands; Amsterdam Institute for

Molecules Medicines and Systems (AIMMS), VU University Amsterdam, Amsterdam, The

Netherlands.

Multiple Sequence Alignment (MSA) methods are typically benchmarked on sets of

reference alignments. The quality of the alignment can then be represented by the

sum-of-pairs (SP) or column (CS) scores, which measure the agreement between a

reference and corresponding query alignment. Both the SP and CS scores treat

mismatches between a query and reference alignment as equally bad, and do not

take the separation into account between two amino acids in the query alignment,

that should have been matched according to the reference alignment. This is

significant since the magnitude of alignment shifts is often of relevance in

biological analyses, including homology modeling and MSA refinement/manual

alignment editing. In this study we develop a new alignment benchmark scoring

scheme, SPdist, that takes the degree of discordance of mismatches into account

by measuring the sequence distance between mismatched residue pairs in the query

alignment. Using this new score along with the standard SP score, we investigate

the discriminatory behavior of the new score by assessing how well six different

MSA methods perform with respect to BAliBASE reference alignments. The SP score

and the SPdist score yield very similar outcomes when the reference and query

alignments are close. However, for more divergent reference alignments the SPdist

score is able to distinguish between methods that keep alignments approximately

close to the reference and those exhibiting larger shifts. We observed that by

using SPdist together with SP scoring we were able to better delineate the

alignment quality difference between alternative MSA methods. With a case study

we exemplify why it is important, from a biological perspective, to consider the

separation of mismatches. The SPdist scoring scheme has been implemented in the

VerAlign web server (http://www.ibi.vu.nl/programs/veralignwww/). The code for

calculating SPdist score is also available upon request.

DOI: 10.1371/journal.pone.0127431

PMCID: PMC4438059

PMID: 25993129 [Indexed for MEDLINE]

569. J Med Chem. 2015 May 14;58(9):4066-72. doi: 10.1021/acs.jmedchem.5b00104. Epub

2015 Apr 22.

pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using

Graph-Based Signatures.

Pires DE(1,)(2), Blundell TL(1), Ascher DB(1).

Author information:

(1)†Department of Biochemistry, University of Cambridge, 80 Tennis Court Road,

Sanger Building, Cambridge, Cambridgshire CB2 1GA, U.K. (2)‡Centro de Pesquisas

René Rachou, Fundação Oswaldo Cruz, Belo Horizonte 30190-002, Brazil.

Drug development has a high attrition rate, with poor pharmacokinetic and safety

properties a significant hurdle. Computational approaches may help minimize these

risks. We have developed a novel approach (pkCSM) which uses graph-based

signatures to develop predictive models of central ADMET properties for drug

development. pkCSM performs as well or better than current methods. A freely

accessible web server (http://structure.bioc.cam.ac.uk/pkcsm), which retains no

information submitted to it, provides an integrated platform to rapidly evaluate

pharmacokinetic and toxicity properties.

DOI: 10.1021/acs.jmedchem.5b00104

PMCID: PMC4434528

PMID: 25860834 [Indexed for MEDLINE]

570. Front Bioeng Biotechnol. 2015 May 8;3:58. doi: 10.3389/fbioe.2015.00058.

eCollection 2015.

SPECTRA: An Integrated Knowledge Base for Comparing Tissue and Tumor-Specific PPI

Networks in Human.

Micale G(1), Ferro A(2), Pulvirenti A(2), Giugno R(2).

Author information:

(1)Department of Computer Science, University of Pisa , Pisa , Italy.

(2)Department of Clinical and Molecular Biomedicine, University of Catania ,

Catania , Italy.

Protein-protein interaction (PPI) networks available in public repositories

usually represent relationships between proteins within the cell. They ignore the

specific set of tissues or tumors where the interactions take place. Indeed,

proteins can form tissue-selective complexes, while they remain inactive in other

tissues. For these reasons, a great attention has been recently paid to

tissue-specific PPI networks, in which nodes are proteins of the global PPI

network whose corresponding genes are preferentially expressed in specific

tissues. In this paper, we present SPECTRA, a knowledge base to build and compare

tissue or tumor-specific PPI networks. SPECTRA integrates gene expression and

protein interaction data from the most authoritative online repositories. We also

provide tools for visualizing and comparing such networks, in order to identify

the expression and interaction changes of proteins across tissues, or between the

normal and pathological states of the same tissue. SPECTRA is available as a web

server at http://alpha.dmi.unict.it/spectra.

DOI: 10.3389/fbioe.2015.00058

PMCID: PMC4424906

PMID: 26005672

571. BMC Res Notes. 2015 May 7;8:187. doi: 10.1186/s13104-015-1152-6.

PR2ALIGN: a stand-alone software program and a web-server for protein sequence

alignment using weighted biochemical properties of amino acids.

Kuznetsov IB(1), McDuffie M(2).

Author information:

(1)Cancer Research Center and Department of Epidemiology and Biostatistics,

University at Albany, State University of New York, One Discovery Drive,

Rensselaer, NY, 12144, USA. ikuznetsov@albany.edu. (2)Cancer Research Center and

Department of Epidemiology and Biostatistics, University at Albany, State

University of New York, One Discovery Drive, Rensselaer, NY, 12144, USA.

mcduffie.michael@gmail.com.

BACKGROUND: Alignment of amino acid sequences is the main sequence comparison

method used in computational molecular biology. The selection of the amino acid

substitution matrix best suitable for a given alignment problem is one of the

most important decisions the user has to make. In a conventional amino acid

substitution matrix all elements are fixed and their values cannot be easily

adjusted. Moreover, most existing amino acid substitution matrices account for

the average (dis)similarities between amino acid types and do not distinguish the

contribution of a specific biochemical property to these (dis)similarities.

FINDINGS: PR2ALIGN is a stand-alone software program and a web-server that

provide the functionality for implementing flexible user-specified alignment

scoring functions and aligning pairs of amino acid sequences based on the

comparison of the profiles of biochemical properties of these sequences. Unlike

the conventional sequence alignment methods that use 20x20 fixed amino acid

substitution matrices, PR2ALIGN uses a set of weighted biochemical properties of

amino acids to measure the distance between pairs of aligned residues and to find

an optimal minimal distance global alignment. The user can provide any number of

amino acid properties and specify a weight for each property. The higher the

weight for a given property, the more this property affects the final alignment.

We show that in many cases the approach implemented in PR2ALIGN produces better

quality pair-wise alignments than the conventional matrix-based approach.

CONCLUSIONS: PR2ALIGN will be helpful for researchers who wish to align amino

acid sequences by using flexible user-specified alignment scoring functions based

on the biochemical properties of amino acids instead of the amino acid

substitution matrix. To the best of the authors' knowledge, there are no existing

stand-alone software programs or web-servers analogous to PR2ALIGN. The software

is freely available from http://pr2align.rit.albany.edu.

DOI: 10.1186/s13104-015-1152-6

PMCID: PMC4477417

PMID: 25947299 [Indexed for MEDLINE]

572. Structure. 2015 May 5;23(5):941-8. doi: 10.1016/j.str.2015.02.013. Epub 2015 Apr

9.

CyToStruct: Augmenting the Network Visualization of Cytoscape with the Power of

Molecular Viewers.

Nepomnyachiy S(1), Ben-Tal N(2), Kolodny R(3).

Author information:

(1)Department of Computer Science & Engineering, Polytechnic Institute of NYU,

Brooklyn, NY 11201, USA. (2)Department of Biochemistry and Molecular

Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat

Aviv 69978, Israel. Electronic address: bental@tauex.tau.ac.il. (3)Department of

Computer Science, University of Haifa, Mount Carmel 31905, Israel. Electronic

address: trachel@cs.haifa.ac.il.

It can be informative to view biological data, e.g., protein-protein interactions

within a large complex, in a network representation coupled with

three-dimensional structural visualizations of individual molecular entities.

CyToStruct, introduced here, provides a transparent interface between the

Cytoscape platform for network analysis and molecular viewers, including PyMOL,

UCSF Chimera, VMD, and Jmol. CyToStruct launches and passes scripts to molecular

viewers from the network's edges and nodes. We provide demonstrations to analyze

interactions among subunits in large protein/RNA/DNA complexes, and similarities

among proteins. CyToStruct enriches the network tools of Cytoscape by adding a

layer of structural analysis, offering all capabilities implemented in molecular

viewers. CyToStruct is available at

https://bitbucket.org/sergeyn/cytostruct/wiki/Home and in the Cytoscape App

Store. Given the coordinates of a molecular complex, our web server

(http://trachel-srv.cs.haifa.ac.il/rachel/ppi/) automatically generates all files

needed to visualize the complex as a Cytoscape network with CyToStruct bridging

to PyMOL, UCSF Chimera, VMD, and Jmol.

Copyright © 2015 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.str.2015.02.013

PMID: 25865247 [Indexed for MEDLINE]

573. Acta Crystallogr D Biol Crystallogr. 2015 May;71(Pt 5):1077-86. doi:

10.1107/S1399004715003144. Epub 2015 Apr 24.

Identification of local variations within secondary structures of proteins.

Kumar P(1), Bansal M(1).

Author information:

(1)Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Karnataka

560 012, India.

Secondary-structure elements (SSEs) play an important role in the folding of

proteins. Identification of SSEs in proteins is a common problem in structural

biology. A new method, ASSP (Assignment of Secondary Structure in Proteins),

using only the path traversed by the C(α) atoms has been developed. The algorithm

is based on the premise that the protein structure can be divided into continuous

or uniform stretches, which can be defined in terms of helical parameters, and

depending on their values the stretches can be classified into different SSEs,

namely α-helices, 310-helices, π-helices, extended β-strands and polyproline II

(PPII) and other left-handed helices. The methodology was validated using an

unbiased clustering of these parameters for a protein data set consisting of 1008

protein chains, which suggested that there are seven well defined clusters

associated with different SSEs. Apart from α-helices and extended β-strands,

310-helices and π-helices were also found to occur in substantial numbers. ASSP

was able to discriminate non-α-helical segments from flanking α-helices, which

were often identified as part of α-helices by other algorithms. ASSP can also

lead to the identification of novel SSEs. It is believed that ASSP could provide

a better understanding of the finer nuances of protein secondary structure and

could make an important contribution to the better understanding of comparatively

less frequently occurring structural motifs. At the same time, it can contribute

to the identification of novel SSEs. A standalone version of the program for the

Linux as well as the Windows operating systems is freely downloadable and a

web-server version is also available at

http://nucleix.mbu.iisc.ernet.in/assp/index.php.

DOI: 10.1107/S1399004715003144

PMID: 25945573 [Indexed for MEDLINE]

574. Ann Pharm Fr. 2015 May;73(3):229-38. doi: 10.1016/j.pharma.2014.09.004. Epub 2014

Nov 11.

[Review and analysis of the evidence on the role and the impact of pharmacists'

activities: Development of an online tool].

[Article in French]

Guérin A(1), Tanguay C(1), Lebel D(1), Bussières JF(2).

Author information:

(1)Département de pharmacie et unité de recherche en pratique pharmaceutique,

centre hospitalier universitaire Sainte-Justine, 3175, chemin de la

Côte-Sainte-Catherine, H3T 1C5 Montréal, Québec, Canada. (2)Département de

pharmacie et unité de recherche en pratique pharmaceutique, centre hospitalier

universitaire Sainte-Justine, 3175, chemin de la Côte-Sainte-Catherine, H3T 1C5

Montréal, Québec, Canada; Faculté de pharmacie, université de Montréal, 2900,

Édouard Montpetit, H3T 1J4 Montréal, Québec, Canada. Electronic address:

jf.bussieres@ssss.gouv.qc.ca.

BACKGROUND: Considering the increase in healthcare expenses, stakeholders need to

make choices, including healthcare program funding, and professional activities

to prioritise.

PURPOSE: The main objective was to list evidences about the role and impact of

pharmacists.

METHODS: Themes were chosen according to three dimensions of the pharmacist

profession: (1) activities, (2) healthcare programs and (3) disorders. A

literature search was conducted for each theme. A bibliographic data sheet was

completed for each article. An analytic data sheet, consisting of descriptive and

impact outcomes, was also completed for the most relevant articles. For each

theme, a synthesis was elaborated. The website Impact Pharmacie

(http://impactpharmacie.org) was developed.

RESULTS: A total of 70 synthesis were written. A total of 1442 articles were

included with a bibliographic data sheet, and 914 with an analytic data sheet.

Six hundred and fifty articles had positive outcomes on the role of the

pharmacist, representing 803 different positive outcome markers. Pharmacists had

positive outcomes on morbidity (n=212), adherence (n=92), costs (n=36), adverse

effects (n=26), drug errors (n=31) and mortality (n=13).

CONCLUSION: This descriptive study presents the review of the evidence on the

role and the impact of pharmacists activities, which led to the Impact Pharmacie

website. This francophone website can contribute to support clinical pharmacy

development, and to a better use of pharmacists in healthcare.

Copyright © 2014 Elsevier Masson SAS. All rights reserved.

DOI: 10.1016/j.pharma.2014.09.004

PMID: 25934531 [Indexed for MEDLINE]

575. Bioinformatics. 2015 May 1;31(9):1487-9. doi: 10.1093/bioinformatics/btu847. Epub

2015 Jan 6.

The anisotropic network model web server at 2015 (ANM 2.0).

Eyal E(1), Lum G(1), Bahar I(1).

Author information:

(1)Cancer Research Institute, Sheba Medical Center, 2 Sheba Rd, Ramat Gan 52621,

Israel and Department of Computational and System Biology, University of

Pittsburgh, 3501 Fifth Ave, Pittsburgh, PA 15213, USA.

SUMMARY: The anisotropic network model (ANM) is one of the simplest yet powerful

tools for exploring protein dynamics. Its main utility is to predict and

visualize the collective motions of large complexes and assemblies near their

equilibrium structures. The ANM server, introduced by us in 2006 helped making

this tool more accessible to non-sophisticated users. We now provide a new

version (ANM 2.0), which allows inclusion of nucleic acids and ligands in the

network model and thus enables the investigation of the collective motions of

protein-DNA/RNA and -ligand systems. The new version offers the flexibility of

defining the system nodes and the interaction types and cutoffs. It also includes

extensive improvements in hardware, software and graphical interfaces.

AVAILABILITY AND IMPLEMENTATION: ANM 2.0 is available at http://anm.csb.pitt.edu

CONTACT: eran.eyal@sheba.health.gov.il, eyal.eran@gmail.com.

© The Author 2015. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu847

PMCID: PMC4410662

PMID: 25568280 [Indexed for MEDLINE]

576. Bioinformatics. 2015 May 1;31(9):1515-8. doi: 10.1093/bioinformatics/btu831. Epub

2014 Dec 24.

The QDREC web server: determining dose-response characteristics of complex

macroparasites in phenotypic drug screens.

Asarnow D(1), Rojo-Arreola L(1), Suzuki BM(2), Caffrey CR(2), Singh R(2).

Author information:

(1)Department of Computer Science, San Francisco State University, San Francisco,

CA, USA, Department of Pathology and Center for Discovery and Innovation in

Parasitic Diseases, University of California, San Francisco, San Francisco, CA,

USA. (2)Department of Computer Science, San Francisco State University, San

Francisco, CA, USA, Department of Pathology and Center for Discovery and

Innovation in Parasitic Diseases, University of California, San Francisco, San

Francisco, CA, USA Department of Computer Science, San Francisco State

University, San Francisco, CA, USA, Department of Pathology and Center for

Discovery and Innovation in Parasitic Diseases, University of California, San

Francisco, San Francisco, CA, USA.

SUMMARY: Neglected tropical diseases (NTDs) caused by helminths constitute some

of the most common infections of the world's poorest people. The etiological

agents are complex and recalcitrant to standard techniques of molecular biology.

Drug screening against helminths has often been phenotypic and typically involves

manual description of drug effect and efficacy. A key challenge is to develop

automated, quantitative approaches to drug screening against helminth diseases.

The quantal dose-response calculator (QDREC) constitutes a significant step in

this direction. It can be used to automatically determine quantitative

dose-response characteristics and half-maximal effective concentration (EC50)

values using image-based readouts from phenotypic screens, thereby allowing

rigorous comparisons of the efficacies of drug compounds. QDREC has been

developed and validated in the context of drug screening for schistosomiasis, one

of the most important NTDs. However, it is equally applicable to general

phenotypic screening involving helminths and other complex parasites.

AVAILABILITY AND IMPLEMENTATION: QDREC is publically available at:

http://haddock4.sfsu.edu/qdrec2/. Source code and datasets are at:

http://tintin.sfsu.edu/projects/phenotypicAssays.html.

CONTACT: rahul@sfsu.edu.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu831

PMCID: PMC4410654

PMID: 25540182 [Indexed for MEDLINE]

577. Bioinformatics. 2015 May 1;31(9):1481-3. doi: 10.1093/bioinformatics/btu837. Epub

2014 Dec 21.

CONSRANK: a server for the analysis, comparison and ranking of docking models

based on inter-residue contacts.

Chermak E(1), Petta A(1), Serra L(1), Vangone A(1), Scarano V(1), Cavallo L(1),

Oliva R(1).

Author information:

(1)Kaust Catalysis Center, King Abdullah University of Science and Technology,

Thuwal, Saudi Arabia, Dipartimento di Informatica ed Applicazioni, Department of

Chemistry and Biology, University of Salerno, Fisciano, SA, Italy and Department

of Sciences and Technologies, University "Parthenope" of Naples, Naples, Italy.

SUMMARY: Herein, we present CONSRANK, a web tool for analyzing, comparing and

ranking protein-protein and protein-nucleic acid docking models, based on the

conservation of inter-residue contacts and its visualization in 2D and 3D

interactive contact maps.

AVAILABILITY AND IMPLEMENTATION: CONSRANK is accessible as a public web tool at

https://www.molnac.unisa.it/BioTools/consrank/.

CONTACT: romina.oliva@uniparthenope.it.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu837

PMID: 25535242 [Indexed for MEDLINE]

578. IEEE/ACM Trans Comput Biol Bioinform. 2015 May-Jun;12(3):611-21. doi:

10.1109/TCBB.2014.2359451.

Disulfide Connectivity Prediction Based on Modelled Protein 3D Structural

Information and Random Forest Regression.

Yu DJ, Li Y, Hu J, Yang X, Yang JY, Shen HB.

Disulfide connectivity is an important protein structural characteristic.

Accurately predicting disulfide connectivity solely from protein sequence helps

to improve the intrinsic understanding of protein structure and function,

especially in the post-genome era where large volume of sequenced proteins

without being functional annotated is quickly accumulated. In this study, a new

feature extracted from the predicted protein 3D structural information is

proposed and integrated with traditional features to form discriminative

features. Based on the extracted features, a random forest regression model is

performed to predict protein disulfide connectivity. We compare the proposed

method with popular existing predictors by performing both cross-validation and

independent validation tests on benchmark datasets. The experimental results

demonstrate the superiority of the proposed method over existing predictors. We

believe the superiority of the proposed method benefits from both the good

discriminative capability of the newly developed features and the powerful

modelling capability of the random forest. The web server implementation, called

TargetDisulfide, and the benchmark datasets are freely available at:

http://csbio.njust.edu.cn/bioinf/TargetDisulfide for academic use.

DOI: 10.1109/TCBB.2014.2359451

PMID: 26357272 [Indexed for MEDLINE]

579. RNA. 2015 May;21(5):1005-17. doi: 10.1261/rna.049346.114. Epub 2015 Mar 24.

SRD: a Staphylococcus regulatory RNA database.

Sassi M(1), Augagneur Y(1), Mauro T(1), Ivain L(1), Chabelskaya S(1), Hallier

M(1), Sallou O(2), Felden B(1).

Author information:

(1)Inserm U835 Biochimie Pharmaceutique, Rennes University, 35043 Rennes, France.

(2)Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes

University, 35043 Rennes, France.

An overflow of regulatory RNAs (sRNAs) was identified in a wide range of

bacteria. We designed and implemented a new resource for the hundreds of sRNAs

identified in Staphylococci, with primary focus on the human pathogen

Staphylococcus aureus. The "Staphylococcal Regulatory RNA Database" (SRD,

http://srd.genouest.org/) compiled all published data in a single interface

including genetic locations, sequences and other features. SRD proposes novel and

simplified identifiers for Staphylococcal regulatory RNAs (srn) based on the

sRNA's genetic location in S. aureus strain N315 which served as a reference.

From a set of 894 sequences and after an in-depth cleaning, SRD provides a list

of 575 srn exempt of redundant sequences. For each sRNA, their experimental

support(s) is provided, allowing the user to individually assess their validity

and significance. RNA-seq analysis performed on strains N315, NCTC8325, and

Newman allowed us to provide further details, upgrade the initial annotation, and

identified 159 RNA-seq independent transcribed sRNAs. The lists of 575 and 159

sRNAs sequences were used to predict the number and location of srns in 18 S.

aureus strains and 10 other Staphylococci. A comparison of the srn contents

within 32 Staphylococcal genomes revealed a poor conservation between species. In

addition, sRNA structure predictions obtained with MFold are accessible. A BLAST

server and the intaRNA program, which is dedicated to target prediction, were

implemented. SRD is the first sRNA database centered on a genus; it is a

user-friendly and scalable device with the possibility to submit new sequences

that should spread in the literature.

© 2015 Sassi et al.; Published by Cold Spring Harbor Laboratory Press for the RNA

Society.

DOI: 10.1261/rna.049346.114

PMCID: PMC4408781

PMID: 25805861 [Indexed for MEDLINE]

580. RNA. 2015 May;21(5):775-85. doi: 10.1261/rna.043612.113. Epub 2015 Mar 20.

miRBoost: boosting support vector machines for microRNA precursor classification.

Tran Vdu T(1), Tempel S(2), Zerath B(3), Zehraoui F(3), Tahi F(4).

Author information:

(1)IBISC - IBGBI, University of Evry, 91037 Evry CEDEX, France Swiss Institute of

Bioinformatics, 1015 Lausanne, Switzerland. (2)IBISC - IBGBI, University of Evry,

91037 Evry CEDEX, France LCB, CNRS UMR 7283, 13009 Marseille, France. (3)IBISC -

IBGBI, University of Evry, 91037 Evry CEDEX, France. (4)IBISC - IBGBI, University

of Evry, 91037 Evry CEDEX, France tahi@ibisc.univ-evry.fr.

Identification of microRNAs (miRNAs) is an important step toward understanding

post-transcriptional gene regulation and miRNA-related pathology. Difficulties in

identifying miRNAs through experimental techniques combined with the huge amount

of data from new sequencing technologies have made in silico discrimination of

bona fide miRNA precursors from non-miRNA hairpin-like structures an important

topic in bioinformatics. Among various techniques developed for this

classification problem, machine learning approaches have proved to be the most

promising. However these approaches require the use of training data, which is

problematic due to an imbalance in the number of miRNAs (positive data) and

non-miRNAs (negative data), which leads to a degradation of their performance. In

order to address this issue, we present an ensemble method that uses a boosting

technique with support vector machine components to deal with imbalanced training

data. Classification is performed following a feature selection on 187 novel and

existing features. The algorithm, miRBoost, performed better in comparison with

state-of-the-art methods on imbalanced human and cross-species data. It also

showed the highest ability among the tested methods for discovering novel miRNA

precursors. In addition, miRBoost was over 1400 times faster than the second most

accurate tool tested and was significantly faster than most of the other tools.

miRBoost thus provides a good compromise between prediction efficiency and

execution time, making it highly suitable for use in genome-wide miRNA precursor

prediction. The software miRBoost is available on our web server

http://EvryRNA.ibisc.univ-evry.fr.

© 2015 Tran et al.; Published by Cold Spring Harbor Laboratory Press for the RNA

Society.

DOI: 10.1261/rna.043612.113

PMCID: PMC4408786

PMID: 25795417 [Indexed for MEDLINE]

581. BMC Bioinformatics. 2015 Apr 28;16:131. doi: 10.1186/s12859-015-0568-2.

coMET: visualisation of regional epigenome-wide association scan results and DNA

co-methylation patterns.

Martin TC(1), Yet I(2), Tsai PC(3), Bell JT(4).

Author information:

(1)Department of Twin Research and Genetic Epidemiology, St Thomas' Hospital

Campus, King's College London, Westminster Bridge Road, London, UK.

tiphaine.martin@kcl.ac.uk. (2)Department of Twin Research and Genetic

Epidemiology, St Thomas' Hospital Campus, King's College London, Westminster

Bridge Road, London, UK. idil.erte@kcl.ac.uk. (3)Department of Twin Research and

Genetic Epidemiology, St Thomas' Hospital Campus, King's College London,

Westminster Bridge Road, London, UK. pei-chien.tsai@kcl.ac.uk. (4)Department of

Twin Research and Genetic Epidemiology, St Thomas' Hospital Campus, King's

College London, Westminster Bridge Road, London, UK. jordana.bell@kcl.ac.uk.

BACKGROUND: Epigenome-wide association scans (EWAS) are an increasingly powerful

and widely-used approach to assess the role of epigenetic variation in human

complex traits. However, this rapidly emerging field lacks dedicated

visualisation tools that can display features specific to epigenetic datasets.

RESULT: We developed coMET, an R package and online tool for visualisation of

EWAS results in a genomic region of interest. coMET generates a regional plot of

epigenetic-phenotype association results and the estimated DNA methylation

correlation between CpG sites (co-methylation), with further options to visualise

genomic annotations based on ENCODE data, gene tracks, reference CpG-sites, and

user-defined features. The tool can be used to display phenotype association

signals and correlation patterns of microarray or sequencing-based DNA

methylation data, such as Illumina Infinium 450k, WGBS, or MeDIP-seq, as well as

other types of genomic data, such as gene expression profiles. The software is

available as a user-friendly online tool from http://epigen.kcl.ac.uk/comet and

as an R Bioconductor package. Source code, examples, and full documentation are

also available from GitHub.

CONCLUSION: Our new software allows visualisation of EWAS results with functional

genomic annotations and with estimation of co-methylation patterns. coMET is

available to a wide audience as an online tool and R package, and can be a

valuable resource to interpret results in the fast growing field of epigenetics.

The software is designed for epigenetic data, but can also be applied to genomic

and functional genomic datasets in any species.

DOI: 10.1186/s12859-015-0568-2

PMCID: PMC4422463

PMID: 25928765 [Indexed for MEDLINE]

582. BMC Bioinformatics. 2015 Apr 16;16:119. doi: 10.1186/s12859-015-0545-9.

CHEXVIS: a tool for molecular channel extraction and visualization.

Masood TB(1), Sandhya S(2), Chandra N(3), Natarajan V(4,)(5).

Author information:

(1)Department of Computer Science and Automation, Indian Institute of Science,

Bangalore, 560012, India. tbmasood@csa.iisc.ernet.in. (2)Department of

Biochemistry, Indian Institute of Science, Bangalore, 560012, India.

sandhya@mbu.iisc.ernet.in. (3)Department of Biochemistry, Indian Institute of

Science, Bangalore, 560012, India. nchandra@biochem.iisc.ernet.in. (4)Department

of Computer Science and Automation, Indian Institute of Science, Bangalore,

560012, India. vijayn@csa.iisc.ernet.in. (5)Supercomputer Education and Research

Centre, Indian Institute of Science, Bangalore, 560012, India.

vijayn@csa.iisc.ernet.in.

BACKGROUND: Understanding channel structures that lead to active sites or

traverse the molecule is important in the study of molecular functions such as

ion, ligand, and small molecule transport. Efficient methods for extracting,

storing, and analyzing protein channels are required to support such studies.

Further, there is a need for an integrated framework that supports computation of

the channels, interactive exploration of their structure, and detailed visual

analysis of their properties.

RESULTS: We describe a method for molecular channel extraction based on the alpha

complex representation. The method computes geometrically feasible channels,

stores both the volume occupied by the channel and its centerline in a unified

representation, and reports significant channels. The representation also

supports efficient computation of channel profiles that help understand channel

properties. We describe methods for effective visualization of the channels and

their profiles. These methods and the visual analysis framework are implemented

in a software tool, CHEXVIS. We apply the method on a number of known channel

containing proteins to extract pore features. Results from these experiments on

several proteins show that CHEXVIS performance is comparable to, and in some

cases, better than existing channel extraction techniques. Using several case

studies, we demonstrate how CHEXVIS can be used to study channels, extract their

properties and gain insights into molecular function.

CONCLUSION: CHEXVIS supports the visual exploration of multiple channels together

with their geometric and physico-chemical properties thereby enabling the

understanding of the basic biology of transport through protein channels. The

CHEXVIS web-server is freely available at http://vgl.serc.iisc.ernet.in/chexvis/

. The web-server is supported on all modern browsers with latest Java plug-in.

DOI: 10.1186/s12859-015-0545-9

PMCID: PMC4411761

PMID: 25888118 [Indexed for MEDLINE]

583. PLoS One. 2015 Apr 16;10(4):e0123261. doi: 10.1371/journal.pone.0123261.

eCollection 2015.

ClusTrack: feature extraction and similarity measures for clustering of

genome-wide data sets.

Rydbeck H(1), Sandve GK(2), Ferkingstad E(3), Simovski B(2), Rye M(4), Hovig

E(5).

Author information:

(1)Department of Informatics, University of Oslo, Oslo, Norway; Department of

Tumour Biology, The Norwegian Radium Hospital, Oslo University Hospital, Oslo,

Norway. (2)Department of Informatics, University of Oslo, Oslo, Norway.

(3)Statistics For Innovation, Norwegian Computing Center, 0314 Oslo, Norway;

Science Institute, University of Iceland, Dunhaga 5, 107 Reykjavik, Iceland.

(4)Department of Cancer Research and Molecular Medicine, Norwegian University of

Science and Technology, Trondheim, Norway. (5)Department of Informatics,

University of Oslo, Oslo, Norway; Department of Tumour Biology, The Norwegian

Radium Hospital, Oslo University Hospital, Oslo, Norway; Department of Medical

Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Oslo,

Norway.

Clustering is a popular technique for explorative analysis of data, as it can

reveal subgroupings and similarities between data in an unsupervised manner.

While clustering is routinely applied to gene expression data, there is a lack of

appropriate general methodology for clustering of sequence-level genomic and

epigenomic data, e.g. ChIP-based data. We here introduce a general methodology

for clustering data sets of coordinates relative to a genome assembly, i.e.

genomic tracks. By defining appropriate feature extraction approaches and

similarity measures, we allow biologically meaningful clustering to be performed

for genomic tracks using standard clustering algorithms. An implementation of the

methodology is provided through a tool, ClusTrack, which allows fine-tuned

clustering analyses to be specified through a web-based interface. We apply our

methods to the clustering of occupancy of the H3K4me1 histone modification in

samples from a range of different cell types. The majority of samples form

meaningful subclusters, confirming that the definitions of features and

similarity capture biological, rather than technical, variation between the

genomic tracks. Input data and results are available, and can be reproduced,

through a Galaxy Pages document at

http://hyperbrowser.uio.no/hb/u/hb-superuser/p/clustrack. The clustering

functionality is available as a Galaxy tool, under the menu option "Specialized

analyzis of tracks", and the submenu option "Cluster tracks based on genome level

similarity", at the Genomic HyperBrowser server: http://hyperbrowser.uio.no/hb/.

DOI: 10.1371/journal.pone.0123261

PMCID: PMC4400084

PMID: 25879845 [Indexed for MEDLINE]

584. Bioinformatics. 2015 Apr 15;31(8):1296-7. doi: 10.1093/bioinformatics/btu817.

Epub 2014 Dec 10.

GSDS 2.0: an upgraded gene feature visualization server.

Hu B(1), Jin J(2), Guo AY(2), Zhang H(2), Luo J(2), Gao G(2).

Author information:

(1)State Key Laboratory of Protein and Plant Gene Research, College of Life

Sciences, Center for Bioinformatics, Peking University, Beijing 100871, People's

Republic of China, College of Life Science, Beijing Normal University, Beijing

100875, People's Republic of China and Department of Systems Biology, College of

Life Science and Technology, Huazhong University of Science and Technology, Wuhan

430074, People's Republic of China State Key Laboratory of Protein and Plant Gene

Research, College of Life Sciences, Center for Bioinformatics, Peking University,

Beijing 100871, People's Republic of China, College of Life Science, Beijing

Normal University, Beijing 100875, People's Republic of China and Department of

Systems Biology, College of Life Science and Technology, Huazhong University of

Science and Technology, Wuhan 430074, People's Republic of China. (2)State Key

Laboratory of Protein and Plant Gene Research, College of Life Sciences, Center

for Bioinformatics, Peking University, Beijing 100871, People's Republic of

China, College of Life Science, Beijing Normal University, Beijing 100875,

People's Republic of China and Department of Systems Biology, College of Life

Science and Technology, Huazhong University of Science and Technology, Wuhan

430074, People's Republic of China.

: Visualizing genes' structure and annotated features helps biologists to

investigate their function and evolution intuitively. The Gene Structure Display

Server (GSDS) has been widely used by more than 60 000 users since its first

publication in 2007. Here, we reported the upgraded GSDS 2.0 with a newly

designed interface, supports for more types of annotation features and formats,

as well as an integrated visual editor for editing the generated figure.

Moreover, a user-specified phylogenetic tree can be added to facilitate further

evolutionary analysis. The full source code is also available for

downloading.AVAILABILITY AND IMPLEMENTATION: Web server and source code are

freely available at http://gsds.cbi.pku.edu.cn.

CONTACT: gaog@mail.cbi.pku.edu.cn or gsds@mail.cbi.pku.edu.cn

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu817

PMCID: PMC4393523

PMID: 25504850 [Indexed for MEDLINE]

585. Bioinformatics. 2015 Apr 15;31(8):1331-3. doi: 10.1093/bioinformatics/btu809.

Epub 2014 Dec 6.

ICMA: an integrated cardiac modeling and analysis platform.

Hussan JR(1), Hunter PJ(1), Gladding PA(1), Greenberg N(2), Christie R(1), Wu

A(1), Sorby H(1), Thomas JD(2).

Author information:

(1)Auckland Bioengineering Institute, University of Auckland, Auckland, New

Zealand, Waitemata District Health Board, North Shore Hospital, Auckland 0622,

New Zealand, National Space Biomedical Research Institute, Houston, TX

77030-1402, USA, Cleveland Clinic Foundation, Cleveland, OH 44195, USA and

Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA.

(2)Auckland Bioengineering Institute, University of Auckland, Auckland, New

Zealand, Waitemata District Health Board, North Shore Hospital, Auckland 0622,

New Zealand, National Space Biomedical Research Institute, Houston, TX

77030-1402, USA, Cleveland Clinic Foundation, Cleveland, OH 44195, USA and

Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand,

Waitemata District Health Board, North Shore Hospital, Auckland 0622, New

Zealand, National Space Biomedical Research Institute, Houston, TX 77030-1402,

USA, Cleveland Clinic Foundation, Cleveland, OH 44195, USA and Feinberg School of

Medicine, Northwestern University, Chicago, IL 60611, USA.

ICMA, a software framework to create 3D finite element models of the left

ventricle from cardiac ultrasound or magnetic resonance imaging (MRI) data, has

been made available as an open-source code. The framework is hardware vendor

independent and uses speckle tracking (endocardial border detection) on

ultrasound (MRI) imaging data in the form of DICOM. Standard American Heart

Association segment-based strain analysis can be performed using a browser-based

interface. The speckle tracking, border detection and model fitting methods are

implemented in C++ using open-source tools. They are wrapped as web services and

orchestrated via a JBOSS-based application server.AVAILABILITY AND

IMPLEMENTATION: The source code for ICMA is freely available under MPL 1.1 or GPL

2.0 or LGPL 2.1 license at https://github.com/ABI-Software-Laboratory/ICMA and a

standalone virtual machine at http://goo.gl/M4lJKH for download.

CONTACT: r.jagir@auckland.ac.nz

SUPPLEMENTARY INFORMATION: Supplementary materials are available at

Bioinformatics online.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu809

PMCID: PMC4393521

PMID: 25481009 [Indexed for MEDLINE]

586. Bioinformatics. 2015 Apr 15;31(8):1293-5. doi: 10.1093/bioinformatics/btu803.

Epub 2014 Dec 4.

GenomeCons: a web server for manipulating multiple genome sequence alignments and

their consensus sequences.

Sato T(1), Suyama M(1).

Author information:

(1)Medical Institute of Bioregulation, Kyushu University and CREST, Japan Science

and Technology Agency, 812-8582 Fukuoka, Japan Medical Institute of

Bioregulation, Kyushu University and CREST, Japan Science and Technology Agency,

812-8582 Fukuoka, Japan.

Genome sequence alignments provide valuable information on many aspects of

molecular biological processes. In this study, we developed a web server,

GenomeCons, for manipulating multiple genome sequence alignments and their

consensus sequences for high-throughput genome sequence analyses. This server

facilitates the visual inspection of multiple genome sequence alignments for a

set of genomic intervals at a time. This allows the user to examine how these

sites are evolutionarily conserved over time for their functional importance. The

server also reports consensus sequences for the input genomic intervals, which

can be applied to downstream analyses such as the identification of common motifs

in the regions determined by ChIP-seq experiments.AVAILABILITY AND

IMPLEMENTATION: GenomeCons is freely accessible at

http://bioinfo.sls.kyushu-u.ac.jp/genomecons/

CONTACT: mikita@bioreg.kyushu-u.ac.jp.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu803

PMID: 25480378 [Indexed for MEDLINE]

587. Bioinformatics. 2015 Apr 15;31(8):1313-5. doi: 10.1093/bioinformatics/btu790.

Epub 2014 Nov 27.

PEASE: predicting B-cell epitopes utilizing antibody sequence.

Sela-Culang I(1), Ashkenazi S(1), Peters B(1), Ofran Y(1).

Author information:

(1)The Goodman Faculty of Life Sciences, Nanotechnology Building, Bar-Ilan

University, Ramat-Gan 52900, Israel and Division of Vaccine Discovery, La Jolla

Institute for Allergy and Immunology, La Jolla, CA 92037, USA.

Antibody epitope mapping is a key step in understanding antibody-antigen

recognition and is of particular interest for drug development, diagnostics and

vaccine design. Most computational methods for epitope prediction are based on

properties of the antigen sequence and/or structure, not taking into account the

antibody for which the epitope is predicted. Here, we introduce PEASE, a web

server predicting antibody-specific epitopes, utilizing the sequence of the

antibody. The predictions are provided both at the residue level and as patches

on the antigen structure. The tradeoff between recall and precision can be tuned

by the user, by changing the default parameters. The results are provided as text

and HTML files as well as a graph, and can be viewed on the antigen 3D

structure.AVAILABILITY AND IMPLEMENTATION: PEASE is freely available on the web

at www.ofranlab.org/PEASE.

CONTACT: yanay@ofranlab.org.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu790

PMID: 25432167 [Indexed for MEDLINE]

588. Reprod Biol Endocrinol. 2015 Apr 15;13:31. doi: 10.1186/s12958-015-0029-9.

Infertility etiologies are genetically and clinically linked with other diseases

in single meta-diseases.

Tarín JJ(1), García-Pérez MA(2,)(3), Hamatani T(4), Cano A(5,)(6).

Author information:

(1)Department of Functional Biology and Physical Anthropology, Faculty of

Biological Sciences, University of Valencia, Burjassot, Valencia, 46100, Spain.

tarinjj@uv.es. (2)Department of Genetics, Faculty of Biological Sciences,

University of Valencia, Burjassot, Valencia, 46100, Spain. migarpe@uv.es.

(3)Research Unit-INCLIVA, Hospital Clínico de Valencia, Valencia, 46010, Spain.

migarpe@uv.es. (4)Department of Obstetrics and Gynecology, Keio University School

of Medicine, Tokyo, 160-8582, Japan. toshiohamatani@z3.keio.jp. (5)Department of

Pediatrics, Obstetrics and Gynecology, Faculty of Medicine, University of

Valencia, Valencia, 46010, Spain. Antonio.cano@uv.es. (6)Service of Obstetrics

and Gynecology, University Clinic Hospital, Valencia, 46010, Spain.

Antonio.cano@uv.es.

Comment in

J Urol. 2015 Dec;194(6):1712.

The present review aims to ascertain whether different infertility etiologies

share particular genes and/or molecular pathways with other pathologies and are

associated with distinct and particular risks of later-life morbidity and

mortality. In order to reach this aim, we use two different sources of

information: (1) a public web server named DiseaseConnect (

http://disease-connect.org ) focused on the analysis of common genes and

molecular mechanisms shared by diseases by integrating comprehensive omics and

literature data; and (2) a literature search directed to find clinical comorbid

relationships of infertility etiologies with only those diseases appearing after

infertility is manifested. This literature search is performed because

DiseaseConnect web server does not discriminate between pathologies emerging

before, concomitantly or after infertility is manifested. Data show that

different infertility etiologies not only share particular genes and/or molecular

pathways with other pathologies but they have distinct clinical relationships

with other diseases appearing after infertility is manifested. In particular, (1)

testicular and high-grade prostate cancer in male infertility; (2) non-fatal

stroke and endometrial cancer, and likely non-fatal coronary heart disease and

ovarian cancer in polycystic ovary syndrome; (3) osteoporosis, psychosexual

dysfunction, mood disorders and dementia in premature ovarian failure; (4) breast

and ovarian cancer in carriers of BRCA1/2 mutations in diminished ovarian

reserve; (5) clear cell and endometrioid histologic subtypes of invasive ovarian

cancer, and likely low-grade serous invasive ovarian cancer, melanoma and

non-Hodgkin lymphoma in endometriosis; and (6) endometrial and ovarian cancer in

idiopathic infertility. The present data endorse the principle that the

occurrence of a disease (in our case infertility) is non-random in the population

and suggest that different infertility etiologies are genetically and clinically

linked with other diseases in single meta-diseases. This finding opens new

insights for clinicians and reproductive biologists to treat infertility problems

using a phenomic approach instead of considering infertility as an isolated and

exclusive disease of the reproductive system/hypothalamic-pituitary-gonadal axis.

In agreement with a previous validation analysis of the utility of DiseaseConnect

web server, the present study does not show a univocal correspondence between

common gene expression and clinical comorbid relationship. Further work is needed

to untangle the potential genetic, epigenetic and phenotypic relationships that

may be present among different infertility etiologies, morbid conditions and

physical/cognitive traits.

DOI: 10.1186/s12958-015-0029-9

PMCID: PMC4404574

PMID: 25880215 [Indexed for MEDLINE]

589. PLoS One. 2015 Apr 13;10(4):e0119317. doi: 10.1371/journal.pone.0119317.

eCollection 2015.

T346Hunter: a novel web-based tool for the prediction of type III, type IV and

type VI secretion systems in bacterial genomes.

Martínez-García PM(1), Ramos C(2), Rodríguez-Palenzuela P(3).

Author information:

(1)Área de Genética, Facultad de Ciencias, Instituto de Hortofruticultura

Subtropical y Mediterránea 'La Mayora', Universidad de Málaga, Consejo Superior

de Investigaciones Científicas (IHSM-UMA-CSIC), Málaga, E-29071, Spain; Centro de

Biotecnología y Genómica de Plantas (CBGP), Universidad Politécnica de

Madrid-Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria,

Parque Científico y Tecnológico de la Universidad Politécnica de Madrid, Campus

de Montegancedo, Pozuelo de Alarcón, Madrid, 28223, Spain. (2)Área de Genética,

Facultad de Ciencias, Instituto de Hortofruticultura Subtropical y Mediterránea

'La Mayora', Universidad de Málaga, Consejo Superior de Investigaciones

Científicas (IHSM-UMA-CSIC), Málaga, E-29071, Spain. (3)Centro de Biotecnología y

Genómica de Plantas (CBGP), Universidad Politécnica de Madrid-Instituto Nacional

de Investigación y Tecnología Agraria y Alimentaria, Parque Científico y

Tecnológico de la Universidad Politécnica de Madrid, Campus de Montegancedo,

Pozuelo de Alarcón, Madrid, 28223, Spain; Departamento de Biotecnología, Escuela

Técnica Superior de Ingenieros Agrónomos, Universidad Politécnica de Madrid,

Avenida Complutense 3, Madrid, 28040, Spain.

T346Hunter (Type Three, Four and Six secretion system Hunter) is a web-based tool

for the identification and localisation of type III, type IV and type VI

secretion systems (T3SS, T4SS and T6SS, respectively) clusters in bacterial

genomes. Non-flagellar T3SS (NF-T3SS) and T6SS are complex molecular machines

that deliver effector proteins from bacterial cells into the environment or into

other eukaryotic or prokaryotic cells, with significant implications for

pathogenesis of the strains encoding them. Meanwhile, T4SS is a more functionally

diverse system, which is involved in not only effector translocation but also

conjugation and DNA uptake/release. Development of control strategies against

bacterial-mediated diseases requires genomic identification of the virulence

arsenal of pathogenic bacteria, with T3SS, T4SS and T6SS being major determinants

in this regard. Therefore, computational methods for systematic identification of

these specialised machines are of particular interest. With the aim of

facilitating this task, T346Hunter provides a user-friendly web-based tool for

the prediction of T3SS, T4SS and T6SS clusters in newly sequenced bacterial

genomes. After inspection of the available scientific literature, we constructed

a database of hidden Markov model (HMM) protein profiles and sequences

representing the various components of T3SS, T4SS and T6SS. T346Hunter performs

searches of such a database against user-supplied bacterial sequences and

localises enriched regions in any of these three types of secretion systems.

Moreover, through the T346Hunter server, users can visualise the predicted

clusters obtained for approximately 1700 bacterial chromosomes and plasmids.

T346Hunter offers great help to researchers in advancing their understanding of

the biological mechanisms in which these sophisticated molecular machines are

involved. T346Hunter is freely available at

http://bacterial-virulence-factors.cbgp.upm.es/T346Hunter.

DOI: 10.1371/journal.pone.0119317

PMCID: PMC4395097

PMID: 25867189 [Indexed for MEDLINE]

590. Anal Biochem. 2015 Apr 1;474:69-77. doi: 10.1016/j.ab.2014.12.009. Epub 2015 Jan

14.

iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide

composition.

Liu Z(1), Xiao X(2), Qiu WR(3), Chou KC(4).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China. (2)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen

333403, China; Gordon Life Science Institute, Boston, MA 02478, USA. Electronic

address: xxiao@gordonlifescience.org. (3)Computer Department, Jing-De-Zhen

Ceramic Institute, Jing-De-Zhen 333403, China. Electronic address:

qiuone@163.com. (4)Gordon Life Science Institute, Boston, MA 02478, USA; Center

of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University,

Jeddah 21589, Saudi Arabia.

Predominantly occurring on cytosine, DNA methylation is a process by which cells

can modify their DNAs to change the expression of gene products. It plays very

important roles in life development but also in forming nearly all types of

cancer. Therefore, knowledge of DNA methylation sites is significant for both

basic research and drug development. Given an uncharacterized DNA sequence

containing many cytosine residues, which one can be methylated and which one

cannot? With the avalanche of DNA sequences generated during the postgenomic age,

it is highly desired to develop computational methods for accurately identifying

the methylation sites in DNA. Using the trinucleotide composition, pseudo amino

acid components, and a dataset-optimizing technique, we have developed a new

predictor called "iDNA-Methyl" that has achieved remarkably higher success rates

in identifying the DNA methylation sites than the existing predictors. A

user-friendly web-server for the new predictor has been established at

http://www.jci-bioinfo.cn/iDNA-Methyl, where users can easily get their desired

results. We anticipate that the web-server predictor will become a very useful

high-throughput tool for basic research and drug development and that the novel

approach and technique can also be used to investigate many other DNA-related

problems and genome analysis.

Copyright © 2014 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2014.12.009

PMID: 25596338 [Indexed for MEDLINE]

591. Bioinformatics. 2015 Apr 1;31(7):1147-9. doi: 10.1093/bioinformatics/btu784. Epub

2014 Nov 29.

LigDig: a web server for querying ligand-protein interactions.

Fuller JC(1), Martinez M(2), Henrich S(2), Stank A(2), Richter S(2), Wade RC(3).

Author information:

(1)Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical

Studies (HITS), 69118 Heidelberg, Center for Molecular Biology Heidelberg

University (ZMBH), DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany and

Interdisciplinary Center for Scientific Computing (IWR), 69120 Heidelberg,

Germany Molecular and Cellular Modeling Group, Heidelberg Institute for

Theoretical Studies (HITS), 69118 Heidelberg, Center for Molecular Biology

Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany and

Interdisciplinary Center for Scientific Computing (IWR), 69120 Heidelberg,

Germany. (2)Molecular and Cellular Modeling Group, Heidelberg Institute for

Theoretical Studies (HITS), 69118 Heidelberg, Center for Molecular Biology

Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany and

Interdisciplinary Center for Scientific Computing (IWR), 69120 Heidelberg,

Germany. (3)Molecular and Cellular Modeling Group, Heidelberg Institute for

Theoretical Studies (HITS), 69118 Heidelberg, Center for Molecular Biology

Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany and

Interdisciplinary Center for Scientific Computing (IWR), 69120 Heidelberg,

Germany Molecular and Cellular Modeling Group, Heidelberg Institute for

Theoretical Studies (HITS), 69118 Heidelberg, Center for Molecular Biology

Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany and

Interdisciplinary Center for Scientific Computing (IWR), 69120 Heidelberg,

Germany Molecular and Cellular Modeling Group, Heidelberg Institute for

Theoretical Studies (HITS), 69118 Heidelberg, Center for Molecular Biology

Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany and

Interdisciplinary Center for Scientific Computing (IWR), 69120 Heidelberg,

Germany Molecular and Cellular Modeling Group, Heidelberg Institute for

Theoretical Studies (HITS), 69118 Heidelberg, Center for Molecular Biology

Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany and

Interdisciplinary Center for Scientific Computing (IWR), 69120 Heidelberg,

Germany.

LigDig is a web server designed to answer questions that previously required

several independent queries to diverse data sources. It also performs basic

manipulations and analyses of the structures of protein-ligand complexes. The

LigDig webserver is modular in design and consists of seven tools, which can be

used separately, or via linking the output from one tool to the next, in order to

answer more complex questions. Currently, the tools allow a user to: (i) perform

a free-text compound search, (ii) search for suitable ligands, particularly

inhibitors, of a protein and query their interaction network, (iii) search for

the likely function of a ligand, (iv) perform a batch search for compound

identifiers, (v) find structures of protein-ligand complexes, (vi) compare

three-dimensional structures of ligand binding sites and (vii) prepare coordinate

files of protein-ligand complexes for further calculations.AVAILABILITY AND

IMPLEMENTATION: LigDig makes use of freely available databases, including ChEMBL,

PubChem and SABIO-RK, and software programs, including cytoscape.js, PDB2PQR,

ProBiS and Fconv. LigDig can be used by non-experts in bio- and chemoinformatics.

LigDig is available at: http://mcm.h-its.org/ligdig.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu784

PMCID: PMC4382906

PMID: 25433696 [Indexed for MEDLINE]

592. Bioinformatics. 2015 Apr 1;31(7):999-1006. doi: 10.1093/bioinformatics/btu791.

Epub 2014 Nov 26.

MetaPSICOV: combining coevolution methods for accurate prediction of contacts and

long range hydrogen bonding in proteins.

Jones DT(1), Singh T(1), Kosciolek T(1), Tetchner S(1).

Author information:

(1)Bioinformatics Group, Department of Computer Science, University College

London, London WC1E 6BT, UK.

MOTIVATION: Recent developments of statistical techniques to infer direct

evolutionary couplings between residue pairs have rendered covariation-based

contact prediction a viable means for accurate 3D modelling of proteins, with no

information other than the sequence required. To extend the usefulness of contact

prediction, we have designed a new meta-predictor (MetaPSICOV) which combines

three distinct approaches for inferring covariation signals from multiple

sequence alignments, considers a broad range of other sequence-derived features

and, uniquely, a range of metrics which describe both the local and global

quality of the input multiple sequence alignment. Finally, we use a two-stage

predictor, where the second stage filters the output of the first stage. This

two-stage predictor is additionally evaluated on its ability to accurately

predict the long range network of hydrogen bonds, including correctly assigning

the donor and acceptor residues.

RESULTS: Using the original PSICOV benchmark set of 150 protein families,

MetaPSICOV achieves a mean precision of 0.54 for top-L predicted long range

contacts-around 60% higher than PSICOV, and around 40% better than CCMpred. In de

novo protein structure prediction using FRAGFOLD, MetaPSICOV is able to improve

the TM-scores of models by a median of 0.05 compared with PSICOV. Lastly, for

predicting long range hydrogen bonding, MetaPSICOV-HB achieves a precision of

0.69 for the top-L/10 hydrogen bonds compared with just 0.26 for the baseline

MetaPSICOV.

AVAILABILITY AND IMPLEMENTATION: MetaPSICOV is available as a freely available

web server at http://bioinf.cs.ucl.ac.uk/MetaPSICOV. Raw data (predicted contact

lists and 3D models) and source code can be downloaded from

http://bioinf.cs.ucl.ac.uk/downloads/MetaPSICOV.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu791

PMCID: PMC4382908

PMID: 25431331 [Indexed for MEDLINE]

593. Bioinformatics. 2015 Apr 1;31(7):1120-3. doi: 10.1093/bioinformatics/btu743. Epub

2014 Nov 20.

CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target

sites.

Naito Y(1), Hino K(1), Bono H(1), Ui-Tei K(2).

Author information:

(1)Database Center for Life Science (DBCLS), National Institute of Genetics,

Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka

411-8540, Japan and Department of Biological Sciences, Graduate School of

Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Database Center for Life Science (DBCLS), National Institute of Genetics,

Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka

411-8540, Japan and Department of Biological Sciences, Graduate School of

Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

(2)Database Center for Life Science (DBCLS), National Institute of Genetics,

Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka

411-8540, Japan and Department of Biological Sciences, Graduate School of

Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

CRISPRdirect is a simple and functional web server for selecting rational

CRISPR/Cas targets from an input sequence. The CRISPR/Cas system is a promising

technique for genome engineering which allows target-specific cleavage of genomic

DNA guided by Cas9 nuclease in complex with a guide RNA (gRNA), that

complementarily binds to a ∼ 20 nt targeted sequence. The target sequence

requirements are twofold. First, the 5'-NGG protospacer adjacent motif (PAM)

sequence must be located adjacent to the target sequence. Second, the target

sequence should be specific within the entire genome in order to avoid off-target

editing. CRISPRdirect enables users to easily select rational target sequences

with minimized off-target sites by performing exhaustive searches against genomic

sequences. The server currently incorporates the genomic sequences of human,

mouse, rat, marmoset, pig, chicken, frog, zebrafish, Ciona, fruit fly, silkworm,

Caenorhabditis elegans, Arabidopsis, rice, Sorghum and budding

yeast.AVAILABILITY: Freely available at http://crispr.dbcls.jp/.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu743

PMCID: PMC4382898

PMID: 25414360 [Indexed for MEDLINE]

594. Bioinformatics. 2015 Apr 1;31(7):1136-7. doi: 10.1093/bioinformatics/btu761. Epub

2014 Nov 18.

XenoSite server: a web-available site of metabolism prediction tool.

Matlock MK(1), Hughes TB(1), Swamidass SJ(1).

Author information:

(1)Division of Biology and Biological Sciences and Department of Pathology and

Immunology, Washington University School of Medicine, St. Louis, MO 63130, USA.

Cytochrome P450 enzymes (P450s) are metabolic enzymes that process the majority

of FDA-approved, small-molecule drugs. Understanding how these enzymes modify

molecule structure is key to the development of safe, effective drugs. XenoSite

server is an online implementation of the XenoSite, a recently published

computational model for P450 metabolism. XenoSite predicts which atomic sites of

a molecule--sites of metabolism (SOMs)--are modified by P450s. XenoSite server

accepts input in common chemical file formats including SDF and SMILES and

provides tools for visualizing the likelihood that each atomic site is a site of

metabolism for a variety of important P450s, as well as a flat file download of

SOM predictions.AVAILABILITY AND IMPLEMENTATION: XenoSite server is available at

http://swami.wustl.edu/xenosite.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please email: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu761

PMID: 25411327 [Indexed for MEDLINE]

595. BMJ Support Palliat Care. 2015 Apr;5 Suppl 1:A24. doi:

10.1136/bmjspcare-2015-000906.76.

PA16 Carer proofing: empowering family carers to design an online tool to meet

their information needs.

O'Sullivan D(1), O'Mahoney P(1).

Author information:

(1)The Carers Association, Ireland.

BACKGROUND: Among the general public there are misconceptions, sensitivities and

taboos relating to Palliative Care. Furthermore, many Palliative Care recipients

and Family Carers are managing with little guidance, knowledge or training.

AIM: The All Island Institute for Hospice and Palliative Care (AIIHPC) funded

this research to identify and evaluate community learning needs in the context of

Palliative Care and propose key components of an electronic learning package and

supporting materials for local communities on the island of Ireland.

METHODS: This research emulated a participant action research model which was

developed to create an empowerment intervention for older people to protect

themselves against financial abuse.(1) The model is guided by principles of

authentic participation and collaboration; namely through interviews and focus

groups with care recipients and Family Carers with experience of Palliative Care

as well as with relevant health care professionals. The research also assembled a

working group of Family Carers.

RESULTS: The working group of Family Carers was supported and empowered to

reflect on their personal experience, on the insight from Health Care

Professionals and on creative methods of information delivery. The Working Group

made specific recommendations about the content of an online learning package and

the delivery mechanisms that would best meet the information needs of those new

to Palliative Care.

CONCLUSION: Placing Family Carers at the centre of the research process resulted

in pragmatic recommendations and should result in the design of an effective

online information tool.

REFERENCE: The O'Donnell Empowerment Model, NCPOP, UCD

http://www.ncpop.ie/WEAAD\_2013\_Presentations.

© 2015, Published by the BMJ Publishing Group Limited. For permission to use

(where not already granted under a licence) please go to

http://group.bmj.com/group/rights-licensing/permissions.

DOI: 10.1136/bmjspcare-2015-000906.76

PMID: 25960502

596. Genomics. 2015 Apr;105(4):197-203. doi: 10.1016/j.ygeno.2015.01.005. Epub 2015

Jan 30.

Identification of protein-interacting nucleotides in a RNA sequence using

composition profile of tri-nucleotides.

Panwar B(1), Raghava GP(2).

Author information:

(1)Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39A,

Chandigarh, India. Electronic address: bharat@imtech.res.in. (2)Bioinformatics

Centre, CSIR-Institute of Microbial Technology, Sector 39A, Chandigarh, India.

Electronic address: http://www.imtech.res.in/raghava/

The RNA-protein interactions play a diverse role in the cells, thus

identification of RNA-protein interface is essential for the biologist to

understand their function. In the past, several methods have been developed for

predicting RNA interacting residues in proteins, but limited efforts have been

made for the identification of protein-interacting nucleotides in RNAs. In order

to discriminate protein-interacting and non-interacting nucleotides, we used

various classifiers (NaiveBayes, NaiveBayesMultinomial, BayesNet,

ComplementNaiveBayes, MultilayerPerceptron, J48, SMO, RandomForest, SMO and

SVM(light)) for prediction model development using various features and achieved

highest 83.92% sensitivity, 84.82 specificity, 84.62% accuracy and 0.62 Matthew's

correlation coefficient by SVM(light) based models. We observed that certain

tri-nucleotides like ACA, ACC, AGA, CAC, CCA, GAG, UGA, and UUU preferred in

protein-interaction. All the models have been developed using a non-redundant

dataset and are evaluated using five-fold cross validation technique. A

web-server called RNApin has been developed for the scientific community

(http://crdd.osdd.net/raghava/rnapin/).

Copyright © 2015 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ygeno.2015.01.005

PMID: 25640448 [Indexed for MEDLINE]

597. Glycobiology. 2015 Apr;25(4):394-402. doi: 10.1093/glycob/cwu121. Epub 2014 Nov

4.

De novo design of a trans-β-N-acetylglucosaminidase activity from a GH1

β-glycosidase by mechanism engineering.

André-Miral C(1), Koné FM(1), Solleux C(1), Grandjean C(1), Dion M(1), Tran V(1),

Tellier C(2).

Author information:

(1)Faculté des Sciences et des Techniques, Unité de Fonctionnalité et Ingénierie

des Protéines (UFIP), Université de Nantes, UMR CNRS 6286, 2 rue de la

Houssinière, BP 92208, Cedex 3 F-44322 Nantes, France. (2)Faculté des Sciences et

des Techniques, Unité de Fonctionnalité et Ingénierie des Protéines (UFIP),

Université de Nantes, UMR CNRS 6286, 2 rue de la Houssinière, BP 92208, Cedex 3

F-44322 Nantes, France charles.tellier@univ-nantes.fr.

Glycoside hydrolases are particularly abundant in all areas of metabolism as they

are involved in the degradation of natural polysaccharides and glycoconjugates.

These enzymes are classified into 133 families (CAZy server, http://www.cazy.org)

in which members of each family have a similar structure and catalytic mechanism.

In order to understand better the structure/function relationships of these

enzymes and their evolution and to develop new robust evolved glycosidases, we

undertook to convert a Family 1 thermostable β-glycosidase into an

exo-β-N-acetylglucosaminidase. This latter activity is totally absent in Family

1, while natural β-hexosaminidases belong to CAZy Families 3, 20 and 84. Using

molecular modeling, we first showed that the docking of N-acetyl-d-glucosamine in

the subsite -1 of the β-glycosidase from Thermus thermophilus (TtβGly) suggested

several steric conflicts with active site amino-acids (N163, E338) induced by the

N-acetyl group. Both N163A and N163D-E338G mutations induced significant

N-acetylglucosaminidase activity in TtβGly. The double mutant N163D-E338G was

also active on the bicyclic oxazoline substrate, suggesting that this mutated

enzyme uses a catalytic mechanism involving a substrate-assisted catalysis with a

noncovalent oxazoline intermediate, similar to the N-acetylglucosaminidases from

Families 20 and 84. Furthermore, a very efficient trans-N-acetylglucosaminidase

activity was observed when the double mutant was incubated in the presence of

NAG-oxazoline as a donor and

N-methyl-O-benzyl-N-(β-d-glucopyranosyl)-hydroxylamine as an acceptor. More

generally, this work demonstrates that it is possible to exchange the

specificities and catalytic mechanisms with minimal changes between

phylogenetically distant protein structures.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/glycob/cwu121

PMID: 25378480 [Indexed for MEDLINE]

598. Hum Mutat. 2015 Apr;36(4):419-24. doi: 10.1002/humu.22767.

mit-o-matic: a comprehensive computational pipeline for clinical evaluation of

mitochondrial variations from next-generation sequencing datasets.

Vellarikkal SK(1), Dhiman H, Joshi K, Hasija Y, Sivasubbu S, Scaria V.

Author information:

(1)Genomics and Molecular Medicine, CSIR Institute of Genomics and Integrative

Biology, Delhi, India; Academy of Scientific and Innovative Research (AcSIR),

Anusandhan Bhawan, Delhi, India.

The human mitochondrial genome has been reported to have a very high mutation

rate as compared with the nuclear genome. A large number of mitochondrial

mutations show significant phenotypic association and are involved in a broad

spectrum of diseases. In recent years, there has been a remarkable progress in

the understanding of mitochondrial genetics. The availability of next-generation

sequencing (NGS) technologies have not only reduced sequencing cost by orders of

magnitude but has also provided us good quality mitochondrial genome sequences

with high coverage, thereby enabling decoding of a number of human mitochondrial

diseases. In this study, we report a computational and experimental pipeline to

decipher the human mitochondrial DNA variations and examine them for their

clinical correlation. As a proof of principle, we also present a clinical study

of a patient with Leigh disease and confirmed maternal inheritance of the

causative allele. The pipeline is made available as a user-friendly online tool

to annotate variants and find haplogroup, disease association, and heteroplasmic

sites. The "mit-o-matic" computational pipeline represents a comprehensive

cloud-based tool for clinical evaluation of mitochondrial genomic variations from

NGS datasets. The tool is freely available at

http://genome.igib.res.in/mitomatic/.

© 2015 WILEY PERIODICALS, INC.

DOI: 10.1002/humu.22767

PMID: 25677119 [Indexed for MEDLINE]

599. J Med Genet. 2015 Apr;52(4):275-81. doi: 10.1136/jmedgenet-2014-102656. Epub 2015

Jan 16.

mirTrios: an integrated pipeline for detection of de novo and rare inherited

mutations from trios-based next-generation sequencing.

Li J(1), Jiang Y(2), Wang T(2), Chen H(2), Xie Q(2), Shao Q(2), Ran X(2), Xia

K(3), Sun ZS(4), Wu J(4).

Author information:

(1)Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing,

China Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, China

State Key Laboratory of Medical Genetics, Central South University, Changsha,

China. (2)Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou,

China. (3)State Key Laboratory of Medical Genetics, Central South University,

Changsha, China. (4)Beijing Institutes of Life Science, Chinese Academy of

Sciences, Beijing, China Institute of Genomic Medicine, Wenzhou Medical

University, Wenzhou, China.

OBJECTIVES: Recently, several studies documented that de novo mutations (DNMs)

play important roles in the aetiology of sporadic diseases. Next-generation

sequencing (NGS) enables variant calling at single-base resolution on a

genome-wide scale. However, accurate identification of DNMs from NGS data still

remains a major challenge. We developed mirTrios, a web server, to accurately

detect DNMs and rare inherited mutations from NGS data in sporadic diseases.

METHODS: The expectation-maximisation (EM) model was adopted to accurately

identify DNMs from variant call files of a trio generated by GATK (Genome

Analysis Toolkit). The GATK results, which contain certain basic properties (such

as PL, PRT and PART), are iteratively integrated into the EM model to strike a

threshold for DNMs detection. Training sets of true and false positive DNMs in

the EM model were built from whole genome sequencing data of 64 trios.

RESULTS: With our in-house whole exome sequencing datasets from 20 trios,

mirTrios totally identified 27 DNMs in the coding region, 25 of which (92.6%) are

validated as true positives. In addition, to facilitate the interpretation of

diverse mutations, mirTrios can also be employed in the identification of rare

inherited mutations. Embedded with abundant annotation of DNMs and rare inherited

mutations, mirTrios also supports known diagnostic variants and causative gene

identification, as well as the prioritisation of novel and promising candidate

genes.

CONCLUSIONS: mirTrios provides an intuitive interface for the general geneticist

and clinician, and can be widely used for detection of DNMs and rare inherited

mutations, and annotation in sporadic diseases. mirTrios is freely available at

http://centre.bioinformatics.zj.cn/mirTrios/.

Published by the BMJ Publishing Group Limited. For permission to use (where not

already granted under a licence) please go to

http://group.bmj.com/group/rights-licensing/permissions.

DOI: 10.1136/jmedgenet-2014-102656

PMID: 25596308 [Indexed for MEDLINE]

600. Mol Biosyst. 2015 Apr;11(4):1194-204. doi: 10.1039/c5mb00050e.

miRNA-dis: microRNA precursor identification based on distance structure status

pairs.

Liu B(1), Fang L, Chen J, Liu F, Wang X.

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen,

Guangdong 518055, China. bliu@insun.hit.edu.cn dragoncloudest@gmail.com

chenjunjie@hitsz.edu.cn liufule12@gmail.com wangxl@insun.hit.edu.cn.

MicroRNA precursor identification is an important task in bioinformatics. Support

Vector Machine (SVM) is one of the most effective machine learning methods used

in this field. The performance of SVM-based methods depends on the vector

representations of RNAs. However, the discriminative power of the existing

feature vectors is limited, and many methods lack an interpretable model for

analysis of characteristic sequence features. Prior studies have demonstrated

that sequence or structure order effects were relevant for discrimination, but

little work has explored how to use this kind of information for human

pre-microRNA identification. In this study, in order to incorporate the

structure-order information into the prediction, a method called "miRNA-dis" was

proposed, in which the feature vector was constructed by the occurrence frequency

of the "distance structure status pair" or just the "distance-pair". Rigorous

cross-validations on a much larger and more stringent newly constructed benchmark

dataset showed that the miRNA-dis outperformed some state-of-the-art predictors

in this area. Remarkably, miRNA-dis trained with human data can correctly predict

87.02% of the 4022 pre-miRNAs from 11 different species ranging from animals,

plants and viruses. miRNA-dis would be a useful high throughput tool for

large-scale analysis of microRNA precursors. In addition, the learnt model can be

easily analyzed in terms of discriminative features, and some interesting

patterns were discovered, which could reflect the characteristics of microRNAs. A

user-friendly web-server of miRNA-dis was constructed, which is freely accessible

to the public at the web-site on http://bioinformatics.hitsz.edu.cn/miRNA-dis/.

DOI: 10.1039/c5mb00050e

PMID: 25715848 [Indexed for MEDLINE]

601. Proteins. 2015 Apr;83(4):640-50. doi: 10.1002/prot.24761. Epub 2015 Feb 5.

Inferring the microscopic surface energy of protein-protein interfaces from

mutation data.

Moal IH(1), Dapkūnas J, Fernández-Recio J.

Author information:

(1)Joint BSC-IRB Research Program in Computational Biology, Life Science

Department, Barcelona Supercomputing Center, Barcelona, 08034, Spain.

Mutations at protein-protein recognition sites alter binding strength by altering

the chemical nature of the interacting surfaces. We present a simple surface

energy model, parameterized with empirical ΔΔG values, yielding mean energies of

-48 cal mol(-1) Å(-2) for interactions between hydrophobic surfaces, -51 to -80

cal mol(-1) Å(-2) for surfaces of complementary charge, and 66-83 cal mol(-1)

Å(-2) for electrostatically repelling surfaces, relative to the aqueous phase.

This places the mean energy of hydrophobic surface burial at -24 cal mol(-1)

Å(-2) . Despite neglecting configurational entropy and intramolecular changes,

the model correlates with empirical binding free energies of a functionally

diverse set of rigid-body interactions (r = 0.66). When used to rerank docking

poses, it can place near-native solutions in the top 10 for 37% of the complexes

evaluated, and 82% in the top 100. The method shows that hydrophobic burial is

the driving force for protein association, accounting for 50-95% of the cohesive

energy. The model is available open-source from

http://life.bsc.es/pid/web/surface\_energy/ and via the CCharpPPI web server

http://life.bsc.es/pid/ccharppi/.

© 2015 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24761

PMID: 25586563 [Indexed for MEDLINE]

602. PLoS Comput Biol. 2015 Mar 31;11(3):e1004153. doi: 10.1371/journal.pcbi.1004153.

eCollection 2015.

Large-scale chemical similarity networks for target profiling of compounds

identified in cell-based chemical screens.

Lo YC(1), Senese S(2), Li CM(3), Hu Q(4), Huang Y(3), Damoiseaux R(5), Torres

JZ(6).

Author information:

(1)Department of Chemistry and Biochemistry, University of California, Los

Angeles, Los Angeles, California, United States of America; Program in

Bioengineering, University of California, Los Angeles, Los Angeles, California,

United States of America. (2)Department of Chemistry and Biochemistry, University

of California, Los Angeles, Los Angeles, California, United States of America.

(3)Drug Studies Unit, Department of Bioengineering & Therapeutic Sciences,

University of California, San Francisco, San Francisco, California, United States

of America. (4)Institute for Digital Research and Education, University of

California, Los Angeles, Los Angeles, California, United States of America.

(5)California NanoSystems Institute, University of California, Los Angeles, Los

Angeles, California, United States of America. (6)Department of Chemistry and

Biochemistry, University of California, Los Angeles, Los Angeles, California,

United States of America; Jonsson Comprehensive Cancer Center, University of

California, Los Angeles, Los Angeles, California, United States of America;

Molecular Biology Institute, University of California, Los Angeles, Los Angeles,

California, United States of America.

Target identification is one of the most critical steps following cell-based

phenotypic chemical screens aimed at identifying compounds with potential uses in

cell biology and for developing novel disease therapies. Current in silico target

identification methods, including chemical similarity database searches, are

limited to single or sequential ligand analysis that have limited capabilities

for accurate deconvolution of a large number of compounds with diverse chemical

structures. Here, we present CSNAP (Chemical Similarity Network Analysis

Pulldown), a new computational target identification method that utilizes

chemical similarity networks for large-scale chemotype (consensus chemical

pattern) recognition and drug target profiling. Our benchmark study showed that

CSNAP can achieve an overall higher accuracy (>80%) of target prediction with

respect to representative chemotypes in large (>200) compound sets, in comparison

to the SEA approach (60-70%). Additionally, CSNAP is capable of integrating with

biological knowledge-based databases (Uniprot, GO) and high-throughput biology

platforms (proteomic, genetic, etc) for system-wise drug target validation. To

demonstrate the utility of the CSNAP approach, we combined CSNAP's target

prediction with experimental ligand evaluation to identify the major mitotic

targets of hit compounds from a cell-based chemical screen and we highlight novel

compounds targeting microtubules, an important cancer therapeutic target. The

CSNAP method is freely available and can be accessed from the CSNAP web server

(http://services.mbi.ucla.edu/CSNAP/).

DOI: 10.1371/journal.pcbi.1004153

PMCID: PMC4380459

PMID: 25826798 [Indexed for MEDLINE]

603. BMC Bioinformatics. 2015 Mar 26;16:101. doi: 10.1186/s12859-015-0531-2.

ProteinVolume: calculating molecular van der Waals and void volumes in proteins.

Chen CR(1), Makhatadze GI(2).

Author information:

(1)Department of Biological Sciences and Center for Biotechnology and

Interdisciplinary Studies, Rensselaer Polytechnic Institute, 110 8th Street,

Troy, NY, 12180, USA. chenc13@rpi.edu. (2)Department of Biological Sciences and

Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic

Institute, 110 8th Street, Troy, NY, 12180, USA. makhag@rpi.edu.

BACKGROUND: Voids and cavities in the native protein structure determine the

pressure unfolding of proteins. In addition, the volume changes due to the

interaction of newly exposed atoms with solvent upon protein unfolding also

contribute to the pressure unfolding of proteins. Quantitative understanding of

these effects is important for predicting and designing proteins with predefined

response to changes in hydrostatic pressure using computational approaches. The

molecular surface volume is a useful metric that describes contribution of

geometrical volume, which includes van der Waals volume and volume of the voids,

to the total volume of a protein in solution, thus isolating the effects of

hydration for separate calculations.

RESULTS: We developed ProteinVolume, a highly robust and easy-to-use tool to

compute geometric volumes of proteins. ProteinVolume generates the molecular

surface of a protein and uses an innovative flood-fill algorithm to calculate the

individual components of the molecular surface volume, van der Waals and

intramolecular void volumes. ProteinVolume is user friendly and is available as a

web-server or a platform-independent command-line version.

CONCLUSIONS: ProteinVolume is a highly accurate and fast application to

interrogate geometric volumes of proteins. ProteinVolume is a free web server

available on http://gmlab.bio.rpi.edu . Free-standing platform-independent

Java-based ProteinVolume executable is also freely available at this web site.

DOI: 10.1186/s12859-015-0531-2

PMCID: PMC4379742

PMID: 25885484 [Indexed for MEDLINE]

604. Biol Direct. 2015 Mar 25;10:10. doi: 10.1186/s13062-015-0046-9.

QSAR based model for discriminating EGFR inhibitors and non-inhibitors using

Random forest.

Singh H(1), Singh S(2), Singla D(3), Agarwal SM(4), Raghava GP(5).

Author information:

(1)Bioinformatics Center, Institute of Microbial Technology, Sector 39-A,

Chandigarh, India. harinder@imtech.res.in. (2)Bioinformatics Center, Institute of

Microbial Technology, Sector 39-A, Chandigarh, India. sandysingh@imtech.res.in.

(3)Bioinformatics Center, Institute of Microbial Technology, Sector 39-A,

Chandigarh, India. deep\_kumar@imtech.res.in. (4)Institute of Cytology and

Preventive Oncology, Sector 39, Noida, 201301, Uttar Pradesh, India.

smagarwal@yahoo.com. (5)Bioinformatics Center, Institute of Microbial Technology,

Sector 39-A, Chandigarh, India. raghava@imtech.res.in.

BACKGROUND: Epidermal Growth Factor Receptor (EGFR) is a well-characterized

cancer drug target. In the past, several QSAR models have been developed for

predicting inhibition activity of molecules against EGFR. These models are useful

to a limited set of molecules for a particular class like

quinazoline-derivatives. In this study, an attempt has been made to develop

prediction models on a large set of molecules (~3500 molecules) that include

diverse scaffolds like quinazoline, pyrimidine, quinoline and indole.

RESULTS: We train, test and validate our classification models on a dataset

called EGFR10 that contains 508 inhibitors (having inhibition activity IC50 less

than 10 nM) and 2997 non-inhibitors. Our Random forest based model achieved

maximum MCC 0.49 with accuracy 83.7% on a validation set using 881 PubChem

fingerprints. In this study, frequency-based feature selection technique has been

used to identify best fingerprints. It was observed that PubChem fingerprints

FP380 (C(~O) (~O)), FP579 (O = C-C-C-C), FP388 (C(:C) (:N) (:N)) and FP 816

(ClC1CC(Br)CCC1) are more frequent in the inhibitors in comparison to

non-inhibitors. In addition, we created different datasets namely EGFR100

containing inhibitors having IC50 < 100 nM and EGFR1000 containing inhibitors

having IC50 < 1000 nM. We trained, test and validate our models on datasets

EGFR100 and EGFR1000 datasets and achieved and maximum MCC 0.58 and 0.71

respectively. In addition, models were developed for predicting quinazoline and

pyrimidine based EGFR inhibitors.

CONCLUSIONS: In summary, models have been developed on a large set of molecules

of various classes for discriminating EGFR inhibitors and non-inhibitors. These

highly accurate prediction models can be used to design and discover novel EGFR

inhibitors. In order to provide service to the scientific community, a web

server/standalone EGFRpred also has been developed (

http://crdd.osdd.net/oscadd/egfrpred/ ).

DOI: 10.1186/s13062-015-0046-9

PMCID: PMC4372225

PMID: 25880749 [Indexed for MEDLINE]

605. PLoS One. 2015 Mar 23;10(3):e0119705. doi: 10.1371/journal.pone.0119705.

eCollection 2015.

TBI server: a web server for predicting ion effects in RNA folding.

Zhu Y(1), He Z(2), Chen SJ(2).

Author information:

(1)Department of Physics, Department of Biochemistry, and Informatics Institute,

University of Missouri, Columbia, MO 65211, USA; Department of Physics, Hangzhou

Normal University, Hangzhou, Zhejiang 310036, China. (2)Department of Physics,

Department of Biochemistry, and Informatics Institute, University of Missouri,

Columbia, MO 65211, USA.

BACKGROUND: Metal ions play a critical role in the stabilization of RNA

structures. Therefore, accurate prediction of the ion effects in RNA folding can

have a far-reaching impact on our understanding of RNA structure and function.

Multivalent ions, especially Mg²⁺, are essential for RNA tertiary structure

formation. These ions can possibly become strongly correlated in the close

vicinity of RNA surface. Most of the currently available software packages, which

have widespread success in predicting ion effects in biomolecular systems,

however, do not explicitly account for the ion correlation effect. Therefore, it

is important to develop a software package/web server for the prediction of ion

electrostatics in RNA folding by including ion correlation effects.

RESULTS: The TBI web server http://rna.physics.missouri.edu/tbi\_index.html

provides predictions for the total electrostatic free energy, the different free

energy components, and the mean number and the most probable distributions of the

bound ions. A novel feature of the TBI server is its ability to account for ion

correlation and ion distribution fluctuation effects.

CONCLUSIONS: By accounting for the ion correlation and fluctuation effects, the

TBI server is a unique online tool for computing ion-mediated electrostatic

properties for given RNA structures. The results can provide important data for

in-depth analysis for ion effects in RNA folding including the ion-dependence of

folding stability, ion uptake in the folding process, and the interplay between

the different energetic components.

DOI: 10.1371/journal.pone.0119705

PMCID: PMC4370743

PMID: 25798933 [Indexed for MEDLINE]

606. J Cheminform. 2015 Mar 22;7:10. doi: 10.1186/s13321-015-0061-y. eCollection 2015.

Wikipedia Chemical Structure Explorer: substructure and similarity searching of

molecules from Wikipedia.

Ertl P(1), Patiny L(2), Sander T(3), Rufener C(3), Zasso M(2).

Author information:

(1)Novartis Institutes for BioMedical Research, Novartis Campus, CH-4056 Basel,

Switzerland. (2)Ecole Polytechnique Fédérale de Lausanne (EPFL), Institute of

Chemical Sciences and Engineering (ISIC), 1015 Lausanne, Switzerland. (3)Actelion

Pharmaceuticals Ltd., Gewerbestrasse 16, CH-4123 Allschwil, Switzerland.

BACKGROUND: Wikipedia, the world's largest and most popular encyclopedia is an

indispensable source of chemistry information. It contains among others also

entries for over 15,000 chemicals including metabolites, drugs, agrochemicals and

industrial chemicals. To provide an easy access to this wealth of information we

decided to develop a substructure and similarity search tool for chemical

structures referenced in Wikipedia.

RESULTS: We extracted chemical structures from entries in Wikipedia and

implemented a web system allowing structure and similarity searching on these

data. The whole search as well as visualization system is written in JavaScript

and therefore can run locally within a web page and does not require a central

server. The Wikipedia Chemical Structure Explorer is accessible on-line at

www.cheminfo.org/wikipedia and is available also as an open source project from

GitHub for local installation.

CONCLUSIONS: The web-based Wikipedia Chemical Structure Explorer provides a

useful resource for research as well as for chemical education enabling both

researchers and students easy and user friendly chemistry searching and

identification of relevant information in Wikipedia. The tool can also help to

improve quality of chemical entries in Wikipedia by providing potential

contributors regularly updated list of entries with problematic structures. And

last but not least this search system is a nice example of how the modern web

technology can be applied in the field of cheminformatics. Graphical

abstractWikipedia Chemical Structure Explorer allows substructure and similarity

searches on molecules referenced in Wikipedia.

DOI: 10.1186/s13321-015-0061-y

PMCID: PMC4374119

PMID: 25815062

607. ACS Synth Biol. 2015 Mar 20;4(3):342-9. doi: 10.1021/sb500194z. Epub 2014 May 21.

Software-supported USER cloning strategies for site-directed mutagenesis and DNA

assembly.

Genee HJ(1), Bonde MT(1), Bagger FO(2,)(3,)(4), Jespersen JB(5,)(6), Sommer

MO(1,)(7), Wernersson R(5,)(8), Olsen LR(2,)(4).

Author information:

(1)†The Novo Nordisk Foundation Center for Biosustainability, Technical

University of Denmark, Hørsholm, Denmark. (2)‡Bioinformatics Centre, Department

of Biology, University of Copenhagen, Copenhagen, Denmark. (3)§The Finsen

Laboratory, Rigshospitalet, Faculty of Health Sciences, Copenhagen, Denmark.

(4)∥Biotech Research and Innovation Center (BRIC), Copenhagen, Denmark.

(5)⊥Center for Biological Sequence Analysis, Department of Systems Biology,

Technical University of Denmark, Lyngby, Denmark. (6)#Department of Chemistry,

Technical University of Denmark, Lyngby, Denmark. (7)∇Department of Systems

Biology, Technical University of Denmark, Lyngby, Denmark. (8)○Intomics A/S,

Lyngby, Denmark.

USER cloning is a fast and versatile method for engineering of plasmid DNA. We

have developed a user friendly Web server tool that automates the design of

optimal PCR primers for several distinct USER cloning-based applications. Our Web

server, named AMUSER (Automated DNA Modifications with USER cloning), facilitates

DNA assembly and introduction of virtually any type of site-directed mutagenesis

by designing optimal PCR primers for the desired genetic changes. To demonstrate

the utility, we designed primers for a simultaneous two-position site-directed

mutagenesis of green fluorescent protein (GFP) to yellow fluorescent protein

(YFP), which in a single step reaction resulted in a 94% cloning efficiency.

AMUSER also supports degenerate nucleotide primers, single insert combinatorial

assembly, and flexible parameters for PCR amplification. AMUSER is freely

available online at http://www.cbs.dtu.dk/services/AMUSER/.

DOI: 10.1021/sb500194z

PMID: 24847672 [Indexed for MEDLINE]

608. PLoS One. 2015 Mar 17;10(3):e0120066. doi: 10.1371/journal.pone.0120066.

eCollection 2015.

Prediction and analysis of quorum sensing peptides based on sequence features.

Rajput A(1), Gupta AK(1), Kumar M(1).

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Council of

Scientific and Industrial Research (CSIR), Sector 39-A, Chandigarh-160036, India.

Quorum sensing peptides (QSPs) are the signaling molecules used by the

Gram-positive bacteria in orchestrating cell-to-cell communication. In spite of

their enormous importance in signaling process, their detailed bioinformatics

analysis is lacking. In this study, QSPs and non-QSPs were examined according to

their amino acid composition, residues position, motifs and physicochemical

properties. Compositional analysis concludes that QSPs are enriched with aromatic

residues like Trp, Tyr and Phe. At the N-terminal, Ser was a dominant residue at

maximum positions, namely, first, second, third and fifth while Phe was a

preferred residue at first, third and fifth positions from the C-terminal. A few

motifs from QSPs were also extracted. Physicochemical properties like

aromaticity, molecular weight and secondary structure were found to be

distinguishing features of QSPs. Exploiting above properties, we have developed a

Support Vector Machine (SVM) based predictive model. During 10-fold

cross-validation, SVM achieves maximum accuracy of 93.00%, Mathew's correlation

coefficient (MCC) of 0.86 and Receiver operating characteristic (ROC) of 0.98 on

the training/testing dataset (T200p+200n). Developed models performed equally

well on the validation dataset (V20p+20n). The server also integrates several

useful analysis tools like "QSMotifScan", "ProtFrag", "MutGen" and "PhysicoProp".

Our analysis reveals important characteristics of QSPs and on the basis of these

unique features, we have developed a prediction algorithm "QSPpred" (freely

available at: http://crdd.osdd.net/servers/qsppred).

DOI: 10.1371/journal.pone.0120066

PMCID: PMC4363368

PMID: 25781990 [Indexed for MEDLINE]

609. Anal Biochem. 2015 Mar 15;473:14-27. doi: 10.1016/j.ab.2014.10.014. Epub 2014 Oct

31.

mPLR-Loc: an adaptive decision multi-label classifier based on penalized logistic

regression for protein subcellular localization prediction.

Wan S(1), Mak MW(2), Kung SY(3).

Author information:

(1)Department of Electronic and Information Engineering, The Hong Kong

Polytechnic University, Hong Kong SAR, China. (2)Department of Electronic and

Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR,

China. Electronic address: enmwmak@polyu.edu.hk. (3)Department of Electrical

Engineering, Princeton University, Princeton, NJ 08544, USA.

Proteins located in appropriate cellular compartments are of paramount importance

to exert their biological functions. Prediction of protein subcellular

localization by computational methods is required in the post-genomic era. Recent

studies have been focusing on predicting not only single-location proteins but

also multi-location proteins. However, most of the existing predictors are far

from effective for tackling the challenges of multi-label proteins. This article

proposes an efficient multi-label predictor, namely mPLR-Loc, based on penalized

logistic regression and adaptive decisions for predicting both single- and

multi-location proteins. Specifically, for each query protein, mPLR-Loc exploits

the information from the Gene Ontology (GO) database by using its accession

number (AC) or the ACs of its homologs obtained via BLAST. The frequencies of GO

occurrences are used to construct feature vectors, which are then classified by

an adaptive decision-based multi-label penalized logistic regression classifier.

Experimental results based on two recent stringent benchmark datasets (virus and

plant) show that mPLR-Loc remarkably outperforms existing state-of-the-art

multi-label predictors. In addition to being able to rapidly and accurately

predict subcellular localization of single- and multi-label proteins, mPLR-Loc

can also provide probabilistic confidence scores for the prediction decisions.

For readers' convenience, the mPLR-Loc server is available online

(http://bioinfo.eie.polyu.edu.hk/mPLRLocServer).

Copyright © 2014 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2014.10.014

PMID: 25449328 [Indexed for MEDLINE]

610. J Comput Chem. 2015 Mar 15;36(7):467-77. doi: 10.1002/jcc.23806. Epub 2014 Dec

15.

STOCK: structure mapper and online coarse-graining kit for molecular simulations.

Bevc S(1), Junghans C, Praprotnik M.

Author information:

(1)Laboratory for Molecular Modeling, National Institute of Chemistry, Hajdrihova

19, Ljubljana, SI-1001, Slovenia.

We present a web toolkit STructure mapper and Online Coarse-graining Kit for

setting up coarse-grained molecular simulations. The kit consists of two tools:

structure mapping and Boltzmann inversion tools. The aim of the first tool is to

define a molecular mapping from high, for example, all-atom, to low, that is,

coarse-grained, resolution. Using a graphical user interface it generates input

files, which are compatible with standard coarse-graining packages, for example,

Versatile Object-oriented Toolkit for Coarse-graining Applications and DL\_CGMAP.

Our second tool generates effective potentials for coarse-grained simulations

preserving the structural properties, for example, radial distribution functions,

of the underlying higher resolution model. The required distribution functions

can be provided by any simulation package. Simulations are performed on a local

machine and only the distributions are uploaded to the server. The applicability

of the toolkit is validated by mapping atomistic pentane and polyalanine

molecules to a coarse-grained representation. Effective potentials are derived

for systems of TIP3P (transferable intermolecular potential 3 point) water

molecules and salt solution. The presented coarse-graining web toolkit is

available at http://stock.cmm.ki.si.

© 2014 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23806

PMID: 25504076

611. J Chem Theory Comput. 2015 Mar 10;11(3):1308-14. doi: 10.1021/ct501085y.

Decoding the Mobility and Time Scales of Protein Loops.

Gu Y(1), Li DW(1), Brüschweiler R(1).

Author information:

(1)Department of Chemistry and Biochemistry and ‡Campus Chemical Instrument

Center, The Ohio State University , Columbus, Ohio 43210, United States.

The flexible nature of protein loops and the time scales of their dynamics are

critical for many biologically important events at the molecular level, such as

protein interaction and recognition processes. In order to obtain a predictive

understanding of the dynamic properties of loops, 500 ns molecular dynamics (MD)

computer simulations of 38 different proteins were performed and validated using

NMR chemical shifts. A total of 169 loops were analyzed and classified into three

types, namely fast loops with correlation times <10 ns, slow loops with

correlation times between 10 and 500 ns, and loops that are static over the

course of the whole trajectory. Chemical and biophysical loop descriptors, such

as amino-acid sequence, average 3D structure, charge distribution,

hydrophobicity, and local contacts were used to develop and parametrize the

ToeLoop algorithm for the prediction of the flexibility and motional time scale

of every protein loop, which is also implemented as a public Web server

(http://spin.ccic.ohio-state.edu/index.php/loop). The results demonstrate that

loop dynamics with their time scales can be predicted rapidly with reasonable

accuracy, which will allow the screening of average protein structures to help

better understand the various roles loops can play in the context of

protein-protein interactions and binding.

DOI: 10.1021/ct501085y

PMID: 26579776 [Indexed for MEDLINE]

612. Sci Rep. 2015 Mar 5;5:8767. doi: 10.1038/srep08767.

Network-assisted genetic dissection of pathogenicity and drug resistance in the

opportunistic human pathogenic fungus Cryptococcus neoformans.

Kim H(1), Jung KW(2), Maeng S(2), Chen YL(3), Shin J(1), Shim JE(1), Hwang S(1),

Janbon G(4), Kim T(5), Heitman J(5), Bahn YS(2), Lee I(1).

Author information:

(1)Department of Biotechnology, College of Life Science and Biotechnology, Yonsei

University, Seoul. 120-749, Korea. (2)Department of Biotechnology, Center for

Fungal Pathogenesis, College of Life Science and Biotechnology, Yonsei

University, Seoul. 120-749, Korea. (3)1] Department of Molecular Genetics and

Microbiology, Medicine, and Pharmacology and Cancer Biology, Duke University

Medical Center, Durham, North Carolina, USA [2] Department of Plant Pathology and

Microbiology, National Taiwan University, Taipei, Taiwan. (4)Institut Pasteur,

Unité Biologie et Pathogénicité Fongiques, Département de Mycologie, F-75015,

Paris, France. (5)Department of Molecular Genetics and Microbiology, Medicine,

and Pharmacology and Cancer Biology, Duke University Medical Center, Durham,

North Carolina, USA.

Cryptococcus neoformans is an opportunistic human pathogenic fungus that causes

meningoencephalitis. Due to the increasing global risk of cryptococcosis and the

emergence of drug-resistant strains, the development of predictive genetics

platforms for the rapid identification of novel genes governing pathogenicity and

drug resistance of C. neoformans is imperative. The analysis of functional

genomics data and genome-scale mutant libraries may facilitate the genetic

dissection of such complex phenotypes but with limited efficiency. Here, we

present a genome-scale co-functional network for C. neoformans, CryptoNet, which

covers ~81% of the coding genome and provides an efficient intermediary between

functional genomics data and reverse-genetics resources for the genetic

dissection of C. neoformans phenotypes. CryptoNet is the first genome-scale

co-functional network for any fungal pathogen. CryptoNet effectively identified

novel genes for pathogenicity and drug resistance using guilt-by-association and

context-associated hub algorithms. CryptoNet is also the first genome-scale

co-functional network for fungi in the basidiomycota phylum, as Saccharomyces

cerevisiae belongs to the ascomycota phylum. CryptoNet may therefore provide

insights into pathway evolution between two distinct phyla of the fungal kingdom.

The CryptoNet web server (www.inetbio.org/cryptonet) is a public resource that

provides an interactive environment of network-assisted predictive genetics for

C. neoformans.

DOI: 10.1038/srep08767

PMCID: PMC4350084

PMID: 25739925 [Indexed for MEDLINE]

613. Acta Crystallogr D Biol Crystallogr. 2015 Mar;71(Pt 3):697-705. doi:

10.1107/S1399004715000383. Epub 2015 Feb 26.

Brickworx builds recurrent RNA and DNA structural motifs into medium- and

low-resolution electron-density maps.

Chojnowski G(1), Waleń T(1), Piątkowski P(1), Potrzebowski W(1), Bujnicki JM(1).

Author information:

(1)International Institute of Molecular and Cell Biology, Trojdena 4, 02-109

Warsaw, Poland.

Brickworx is a computer program that builds crystal structure models of nucleic

acid molecules using recurrent motifs including double-stranded helices. In a

first step, the program searches for electron-density peaks that may correspond

to phosphate groups; it may also take into account phosphate-group positions

provided by the user. Subsequently, comparing the three-dimensional patterns of

the P atoms with a database of nucleic acid fragments, it finds the matching

positions of the double-stranded helical motifs (A-RNA or B-DNA) in the unit

cell. If the target structure is RNA, the helical fragments are further extended

with recurrent RNA motifs from a fragment library that contains single-stranded

segments. Finally, the matched motifs are merged and refined in real space to

find the most likely conformations, including a fit of the sequence to the

electron-density map. The Brickworx program is available for download and as a

web server at http://iimcb.genesilico.pl/brickworx.

DOI: 10.1107/S1399004715000383

PMCID: PMC4356372

PMID: 25760616 [Indexed for MEDLINE]

614. Bioinformatics. 2015 Mar 1;31(5):779-81. doi: 10.1093/bioinformatics/btu718. Epub

2014 Oct 28.

WebPSN: a web server for high-throughput investigation of structural

communication in biomacromolecules.

Seeber M(1), Felline A(1), Raimondi F(1), Mariani S(1), Fanelli F(1).

Author information:

(1)Department of Life Sciences, University of Modena and Reggio Emilia, via Campi

183, 41125 Modena, Italy.

We developed a mixed Protein Structure Network (PSN) and Elastic Network

Model-Normal Mode Analysis (ENM-NMA)-based strategy (i.e. PSN-ENM) to investigate

structural communication in biomacromolecules. The approach starts from a Protein

Structure Graph and searches for all shortest communication pathways between

user-specified residues. The graph is computed on a single preferably

high-resolution structure. Information on system's dynamics is supplied by

ENM-NMA. The PSN-ENM methodology is made of multiple steps both in the setup and

analysis stages, which may discourage inexperienced users. To facilitate its

usage, we implemented WebPSN, a freely available web server that allows the user

to easily setup the calculation, perform post-processing analyses and both

visualize and download numerical and 3D representations of the output. Speed and

accuracy make this server suitable to investigate structural communication,

including allosterism, in large sets of bio-macromolecular systems.AVAILABILITY

AND IMPLEMENTATION: The WebPSN server is freely available at

http://webpsn.hpc.unimore.it.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu718

PMID: 25355786 [Indexed for MEDLINE]

615. Bioinformatics. 2015 Mar 1;31(5):764-6. doi: 10.1093/bioinformatics/btu712. Epub

2014 Oct 27.

Navigating protected genomics data with UCSC Genome Browser in a Box.

Haeussler M(1), Raney BJ(1), Hinrichs AS(1), Clawson H(1), Zweig AS(1), Karolchik

D(1), Casper J(1), Speir ML(1), Haussler D(2), Kent WJ(1).

Author information:

(1)Center for Biomolecular Science and Engineering, School of Engineering,

University of California Santa Cruz, Santa Cruz, CA 95064, USA and Howard Hughes

Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064,

USA. (2)Center for Biomolecular Science and Engineering, School of Engineering,

University of California Santa Cruz, Santa Cruz, CA 95064, USA and Howard Hughes

Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA

Center for Biomolecular Science and Engineering, School of Engineering,

University of California Santa Cruz, Santa Cruz, CA 95064, USA and Howard Hughes

Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064,

USA.

Genome Browser in a Box (GBiB) is a small virtual machine version of the popular

University of California Santa Cruz (UCSC) Genome Browser that can be run on a

researcher's own computer. Once GBiB is installed, a standard web browser is used

to access the virtual server and add personal data files from the local hard

disk. Annotation data are loaded on demand through the Internet from UCSC or can

be downloaded to the local computer for faster access.AVAILABILITY AND

IMPLEMENTATION: Software downloads and installation instructions are freely

available for non-commercial use at https://genome-store.ucsc.edu/. GBiB requires

the installation of open-source software VirtualBox, available for all major

operating systems, and the UCSC Genome Browser, which is open source and free for

non-commercial use. Commercial use of GBiB and the Genome Browser requires a

license (http://genome.ucsc.edu/license/).

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu712

PMCID: PMC4341066

PMID: 25348212 [Indexed for MEDLINE]

616. Bioinformatics. 2015 Mar 1;31(5):773-5. doi: 10.1093/bioinformatics/btu709. Epub

2014 Oct 27.

SEABED: Small molEcule activity scanner weB servicE baseD.

Fenollosa C(1), Otón M(1), Andrio P(2), Cortés J(1), Orozco M(3), Goñi JR(1).

Author information:

(1)Department of Life Sciences, Barcelona Supercomputing Center, Barcelona 08034,

Spain, Joint BSC-CRG-IRB Program in Computational Biology, Barcelona, Spain,

Computational Bioinformatics Node, National Institute of Bioinformatics,

Barcelona 08034, Spain, Institute for Research in Biomedicine (IRB Barcelona),

Barcelona 08028, Spain, Department of Biochemistry and Molecular Biology, Biology

Faculty, University of Barcelona, Barcelona 08028, Spain and Structural

Bioinformatics Node, National Institute of Bioinformatics, Barcelona 08028, Spain

Department of Life Sciences, Barcelona Supercomputing Center, Barcelona 08034,

Spain, Joint BSC-CRG-IRB Program in Computational Biology, Barcelona, Spain,

Computational Bioinformatics Node, National Institute of Bioinformatics,

Barcelona 08034, Spain, Institute for Research in Biomedicine (IRB Barcelona),

Barcelona 08028, Spain, Department of Biochemistry and Molecular Biology, Biology

Faculty, University of Barcelona, Barcelona 08028, Spain and Structural

Bioinformatics Node, National Institute of Bioinformatics, Barcelona 08028,

Spain. (2)Department of Life Sciences, Barcelona Supercomputing Center, Barcelona

08034, Spain, Joint BSC-CRG-IRB Program in Computational Biology, Barcelona,

Spain, Computational Bioinformatics Node, National Institute of Bioinformatics,

Barcelona 08034, Spain, Institute for Research in Biomedicine (IRB Barcelona),

Barcelona 08028, Spain, Department of Biochemistry and Molecular Biology, Biology

Faculty, University of Barcelona, Barcelona 08028, Spain and Structural

Bioinformatics Node, National Institute of Bioinformatics, Barcelona 08028, Spain

Department of Life Sciences, Barcelona Supercomputing Center, Barcelona 08034,

Spain, Joint BSC-CRG-IRB Program in Computational Biology, Barcelona, Spain,

Computational Bioinformatics Node, National Institute of Bioinformatics,

Barcelona 08034, Spain, Institute for Research in Biomedicine (IRB Barcelona),

Barcelona 08028, Spain, Department of Biochemistry and Molecular Biology, Biology

Faculty, University of Barcelona, Barcelona 08028, Spain and Structural

Bioinformatics Node, National Institute of Bioinformatics, Barcelona 08028, Spain

Department of Life Sciences, Barcelona Supercomputing Center, Barcelona 08034,

Spain, Joint BSC-CRG-IRB Program in Computational Biology, Barcelona, Spain,

Computational Bioinformatics Node, National Institute of Bioinformatics,

Barcelona 08034, Spain, Institute for Research in Biomedicine (IRB Barcelona),

Barcelona 08028, Spain, Department of Biochemistry and Molecular Biology, Biology

Faculty, University of Barcelona, Barcelona 08028, Spain and Structural

Bioinformatics Node, National Institute of Bioinformatics, Barcelona 08028,

Spain. (3)Department of Life Sciences, Barcelona Supercomputing Center, Barcelona

08034, Spain, Joint BSC-CRG-IRB Program in Computational Biology, Barcelona,

Spain, Computational Bioinformatics Node, National Institute of Bioinformatics,

Barcelona 08034, Spain, Institute for Research in Biomedicine (IRB Barcelona),

Barcelona 08028, Spain, Department of Biochemistry and Molecular Biology, Biology

Faculty, University of Barcelona, Barcelona 08028, Spain and Structural

Bioinformatics Node, National Institute of Bioinformatics, Barcelona 08028, Spain

Department of Life Sciences, Barcelona Supercomputing Center, Barcelona 08034,

Spain, Joint BSC-CRG-IRB Program in Computational Biology, Barcelona, Spain,

Computational Bioinformatics Node, National Institute of Bioinformatics,

Barcelona 08034, Spain, Institute for Research in Biomedicine (IRB Barcelona),

Barcelona 08028, Spain, Department of Biochemistry and Molecular Biology, Biology

Faculty, University of Barcelona, Barcelona 08028, Spain and Structural

Bioinformatics Node, National Institute of Bioinformatics, Barcelona 08028, Spain

Department of Life Sciences, Barcelona Supercomputing Center, Barcelona 08034,

Spain, Joint BSC-CRG-IRB Program in Computational Biology, Barcelona, Spain,

Computational Bioinformatics Node, National Institute of Bioinformatics,

Barcelona 08034, Spain, Institute for Research in Biomedicine (IRB Barcelona),

Barcelona 08028, Spain, Department of Biochemistry and Molecular Biology, Biology

Faculty, University of Barcelona, Barcelona 08028, Spain and Structural

Bioinformatics Node, National Institute of Bioinformatics, Barcelona 08028, Spain

Department of Life Sciences, Barcelona Supercomputing Center, Barcelona 08034,

Spain, Joint BSC-CRG-IRB Program in Computational Biology, Barcelona, Spain,

Computational Bioinformatics Node, National Institute of Bioinformatics,

Barcelona 08034, Spain, Institute for Research in Biomedicine (IRB Barcelona),

Barcelona 08028, Spain, Department of Biochemistry and Molec

MOTIVATION: The SEABED web server integrates a variety of docking and QSAR

techniques in a user-friendly environment. SEABED goes beyond the basic docking

and QSAR web tools and implements extended functionalities like receptor

preparation, library editing, flexible ensemble docking, hybrid docking/QSAR

experiments or virtual screening on protein mutants. SEABED is not a monolithic

workflow tool but Software as a Service platform.

AVAILABILITY AND IMPLEMENTATION: SEABED is a free web server available at

http://www.bsc.es/SEABED.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu709

PMID: 25348211 [Indexed for MEDLINE]

617. Biotechniques. 2015 Mar 1;58(3):140-2. doi: 10.2144/000114266. eCollection 2015.

SIFTER-T: a scalable and optimized framework for the SIFTER phylogenomic method

of probabilistic protein domain annotation.

Almeida-e-Silva DC(1), Vêncio RZ(1).

Author information:

(1)Department of Computing and Mathematics FFCLRP-USP, University of Sao Paulo,

Ribeirão Preto, Brazil.

Statistical Inference of Function Through Evolutionary Relationships (SIFTER) is

a powerful computational platform for probabilistic protein domain annotation.

Nevertheless, SIFTER is not widely used, likely due to usability and scalability

issues. Here we present SIFTER-T (SIFTER Throughput-optimized), a substantial

improvement over SIFTER's original proof-of-principle implementation. SIFTER-T is

optimized for better performance, allowing it to be used at the genome-wide

scale. Compared to SIFTER 2.0, SIFTER-T achieved an 87-fold performance

improvement using published test data sets for the known annotations recovering

module and a 72.3% speed increase for the gene tree generation module in

quad-core machines, as well as a major decrease in memory usage during the

realignment phase. Memory optimization allowed an expanded set of proteins to be

handled by SIFTER's probabilistic method. The improvement in performance and

automation that we achieved allowed us to build a web server to bring the power

of Bayesian phylogenomic inference to the genomics community. SIFTER-T and its

online interface are freely available under GNU license at

http://labpib.fmrp.usp.br/methods/SIFTER-t/ and

https://github.com/dcasbioinfo/SIFTER-t.

DOI: 10.2144/000114266

PMID: 25757547 [Indexed for MEDLINE]

618. Mol Cell Proteomics. 2015 Mar;14(3):761-70. doi: 10.1074/mcp.M114.037994. Epub

2015 Jan 20.

Systematic characterization and prediction of post-translational modification

cross-talk.

Huang Y(1), Xu B(2), Zhou X(3), Li Y(4), Lu M(4), Jiang R(5), Li T(6).

Author information:

(1)From the ‡Department of Biomedical Informatics, ‖MOE Key Laboratory of

Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation,

Tsinghua University, Beijing 100084, China; \*\*European Molecular Biology

Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus,

Hinxton, Cambridge, United Kingdom; (2)§Department of Biochemistry and Molecular

Biology, and. (3)¶¶Department of Psychiatry and Centre for Genomic Sciences, Li

Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China.

(4)From the ‡Department of Biomedical Informatics. (5)‖MOE Key Laboratory of

Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation,

Tsinghua University, Beijing 100084, China; (6)From the ‡Department of Biomedical

Informatics, ¶Institute of Systems Biomedicine, School of Basic Medical Sciences,

Peking University Health Science Center, Beijing 100191, China;

litt@hsc.pku.edu.cn.

Post-translational modification (PTM)(1) plays an important role in regulating

the functions of proteins. PTMs of multiple residues on one protein may work

together to determine a functional outcome, which is known as PTM cross-talk.

Identification of PTM cross-talks is an emerging theme in proteomics and has

elicited great interest, but their properties remain to be systematically

characterized. To this end, we collected 193 PTM cross-talk pairs in 77 human

proteins from the literature and then tested location preference and co-evolution

at the residue and modification levels. We found that cross-talk events

preferentially occurred among nearby PTM sites, especially in disordered protein

regions, and cross-talk pairs tended to co-evolve. Given the properties of PTM

cross-talk pairs, a naïve Bayes classifier integrating different features was

built to predict cross-talks for pairwise combination of PTM sites. By using a

10-fold cross-validation, the integrated prediction model showed an area under

the receiver operating characteristic (ROC) curve of 0.833, superior to using any

individual feature alone. The prediction performance was also demonstrated to be

robust to the biases in the collected PTM cross-talk pairs. The integrated

approach has the potential for large-scale prioritization of PTM cross-talk

candidates for functional validation and was implemented as a web server

available at http://bioinfo.bjmu.edu.cn/ptm-x/.

© 2015 by The American Society for Biochemistry and Molecular Biology, Inc.

DOI: 10.1074/mcp.M114.037994

PMCID: PMC4349993

PMID: 25605461 [Indexed for MEDLINE]

619. Pract Radiat Oncol. 2015 Mar-Apr;5(2):127-34. doi: 10.1016/j.prro.2014.10.012.

Epub 2014 Nov 3.

Safety Profile Assessment: An online tool to gauge safety-critical performance in

radiation oncology.

Dunscombe P(1), Brown D(2), Donaldson H(3), Greener A(4), O'Neill M(5), Sutlief

S(2), Woodward M(6), Yorke E(7), Ford E(8).

Author information:

(1)Department of Oncology, University of Calgary, Calgary, Alberta, Canada.

(2)Radiation Medicine and Applied Sciences, University of California San Diego,

La Jolla, California. (3)Australasian College of Emergency Medicine, Melbourne,

Victoria, Australia. (4)Veterans Affairs New Jersey Health Care System, East

Orange Campus, Radiation Oncology Service, East Orange, New Jersey. (5)Lynchburg

Radiation Oncology Associates, Lynchburg, Virginia. (6)American Association of

Physicists in Medicine, College Park, Maryland. (7)Medical Physics Department,

Memorial Sloan Kettering Cancer Center, New York, New York. (8)Department of

Radiation Oncology, University of Washington, Seattle, Washington. Electronic

address: eford@uw.edu.

PURPOSE: It is challenging for the radiation oncology practitioner to manage and

implement the plethora of recently generated recommendations on quality and

safety improvement. The online Safety Profile Assessment (SPA) tool uses an

easy-to-use question-and-answer format to assess safety/quality within a clinic,

provide a way to benchmark against peers, and facilitate improvement. This report

describes the design and development of the SPA and experience from the first

year of use.

METHODS: Performance indicators for the SPA were derived from 4 foundations: the

Agency for Healthcare Research and Quality, a review of 7 recent authoritative

documents specific to radiation oncology, a recent American Association of

Physicists in Medicine report on incident learning, and the American College of

Radiology-American Society for Radiation Oncology accreditation system as of

2011. After pilot testing, the free-access tool was launched through the American

Association of Physicists in Medicine website (http://spa.aapm.org) in July 2013.

Questionnaire data were collected to assess the experience of users.

RESULTS: The SPA tool consists of 92 indicators designed to probe safety and

quality. A clinic's performance is benchmarked against all other responses in the

database, and aided by a downloadable log, quality/safety improvement strategies

can be developed and tracked over time. At the time this paper was written, 279

individuals had registered, and 107 had completed the SPA. On average, the SPA

required 1.3 hours to complete. The majority of respondents to the questionnaire

(56%) completed the SPA with a multidisciplinary group of 4 people on average.

Respondents noted that the SPA was easy or very easy to use (70%) and that they

would definitely or very probably complete it again (63%).

CONCLUSIONS: SPA provides a straightforward means of gauging a clinic's

performance in key safety-critical areas and has been evaluated favorably by the

first cohort of users. The tool has been qualified by the American Board of

Radiology (ABR) as meeting the criteria for Practice Quality Improvement

requirements of the ABR Maintenance of Certification Program.

Copyright © 2015 The Authors. Published by Elsevier Inc. All rights reserved.

DOI: 10.1016/j.prro.2014.10.012

PMID: 25748005 [Indexed for MEDLINE]

620. Proteins. 2015 Mar;83(3):411-27. doi: 10.1002/prot.24746. Epub 2015 Jan 13.

Refinement by shifting secondary structure elements improves sequence alignments.

Tong J(1), Pei J, Otwinowski Z, Grishin NV.

Author information:

(1)Department of Biophysics, University of Texas Southwestern Medical Center at

Dallas, Dallas, Texas, 75390; Department of Biochemistry, University of Texas

Southwestern Medical Center at Dallas, Dallas, Texas, 75390.

Constructing a model of a query protein based on its alignment to a homolog with

experimentally determined spatial structure (the template) is still the most

reliable approach to structure prediction. Alignment errors are the main

bottleneck for homology modeling when the query is distantly related to the

template. Alignment methods often misalign secondary structural elements by a few

residues. Therefore, better alignment solutions can be found within a limited set

of local shifts of secondary structures. We present a refinement method to

improve pairwise sequence alignments by evaluating alignment variants generated

by local shifts of template-defined secondary structures. Our method SFESA is

based on a novel scoring function that combines the profile-based sequence score

and the structure score derived from residue contacts in a template. Such a

combined score frequently selects a better alignment variant among a set of

candidate alignments generated by local shifts and leads to overall increase in

alignment accuracy. Evaluation of several benchmarks shows that our refinement

method significantly improves alignments made by automatic methods such as

PROMALS, HHpred and CNFpred. The web server is available at

http://prodata.swmed.edu/sfesa.

© 2014 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24746

PMCID: PMC4501258

PMID: 25546158 [Indexed for MEDLINE]

621. BMC Genomics. 2015 Feb 28;16:142. doi: 10.1186/s12864-015-1345-3.

ColoWeb: a resource for analysis of colocalization of genomic features.

Kim R(1), Smith OK(2), Wong WC(3), Ryan AM(4), Ryan MC(5), Aladjem MI(6).

Author information:

(1)In Silico Solutions, Falls Church, VA, USA. rkim@insilico.us.com. (2)DNA

Replication Group, Laboratory of Molecular Pharmacology, Developmental

Therapeutics Branch, National Cancer Institute, National Institutes of Health,

Bethesda, Maryland, USA. owen.smith@nih.gov. (3)In Silico Solutions, Falls

Church, VA, USA. awong@insilico.us.com. (4)In Silico Solutions, Falls Church, VA,

USA. aryan@insilico.us.com. (5)In Silico Solutions, Falls Church, VA, USA.

mryan@insilico.us.com. (6)DNA Replication Group, Laboratory of Molecular

Pharmacology, Developmental Therapeutics Branch, National Cancer Institute,

National Institutes of Health, Bethesda, Maryland, USA. aladjemm@mail.nih.gov.

BACKGROUND: Next-generation sequencing techniques such as ChIP-seq allow

researchers to investigate the genomic position of nuclear components and events.

These experiments provide researchers with thousands of regions of interest to

probe in order to identify biological relevance. As the cost of sequencing

decreases and its robustness increases, more and more researchers turn to

genome-wide studies to better understand the genomic elements they are studying.

One way to interpret the output of sequencing studies is to investigate how the

element of interest localizes in relationship to genome annotations and the

binding of other nuclear components. Colocalization of genomic features could

indicate cooperation and provide evidence for more detailed investigations.

Although there are several existing tools for visualizing and analyzing

colocalization, either they are difficult to use for experimental researchers,

not well maintained, or without measurements for colocalization strength. Here we

describe an online tool, ColoWeb, designed to allow experimentalists to compare

their datasets to existing genomic features in order to generate hypotheses about

biological interactions easily and quickly.

RESULTS: ColoWeb is a web-based service for evaluating the colocation of genomic

features. Users submit genomic regions of interest, for example, a set of

locations from a ChIP-seq analysis. ColoWeb compares the submitted regions of

interest to the location of other genomic features such as transcription factors

and chromatin modifiers. To facilitate comparisons among various genomic

features, the output consists of both graphical representations and quantitative

measures of the degree of colocalization between user's genomic regions and

selected features. Frequent colocation may indicate a biological relationship.

CONCLUSION: ColoWeb is a biologist-friendly web service that can quickly provide

an assessment of thousands of genomic regions to identify colocated genomic

features. ColoWeb is freely available at: http://projects.insilico.us.com/ColoWeb

.

DOI: 10.1186/s12864-015-1345-3

PMCID: PMC4364483

PMID: 25887597 [Indexed for MEDLINE]

622. Database (Oxford). 2015 Feb 27;2015. pii: bav007. doi: 10.1093/database/bav007.

Print 2015.

PreDREM: a database of predicted DNA regulatory motifs from 349 human cell and

tissue samples.

Zheng Y(1), Li X(2), Hu H(3).

Author information:

(1)Department of Electrical Engineering and Computer Science and Burnett School

of Biomedical Science, College of Medicine, University of Central Florida,

Orlando, FL 32816, USA. (2)Department of Electrical Engineering and Computer

Science and Burnett School of Biomedical Science, College of Medicine, University

of Central Florida, Orlando, FL 32816, USA xiaoman@mail.ucf.edu. (3)Department of

Electrical Engineering and Computer Science and Burnett School of Biomedical

Science, College of Medicine, University of Central Florida, Orlando, FL 32816,

USA haihu@cs.ucf.edu.

PreDREM is a database of DNA regulatory motifs and motifs modules predicted from

DNase I hypersensitive sites in 349 human cell and tissue samples. It contains

845-1325 predicted motifs in each sample, which result in a total of 2684

non-redundant motifs. In comparison with seven large collections of known motifs,

more than 84% of the 2684 predicted motifs are similar to the known motifs, and

54-76% of the known motifs are similar to the predicted motifs. PreDREM also

stores 43 663-20 13 288 motif modules in each sample, which provide the cofactor

motifs of each predicted motif. Compared with motifs of known interacting

transcription factor (TF) pairs in eight resources, on average, 84% of motif

pairs corresponding to known interacting TF pairs are included in the predicted

motif modules. Through its web interface, PreDREM allows users to browse motif

information by tissues, datasets, individual non-redundant motifs, etc. Users can

also search motifs, motif modules, instances of motifs and motif modules in given

genomic regions, tissue or cell types a motif occurs, etc. PreDREM thus provides

a useful resource for the understanding of cell- and tissue-specific gene

regulation in the human genome. Database URL: http://server.cs.ucf.edu/predrem/.

© The Author(s) 2015. Published by Oxford University Press.

DOI: 10.1093/database/bav007

PMCID: PMC4343075

PMID: 25725063 [Indexed for MEDLINE]

623. BMC Syst Biol. 2015 Feb 26;9:10. doi: 10.1186/s12918-015-0151-5.

Mimoza: web-based semantic zooming and navigation in metabolic networks.

Zhukova A(1), Sherman DJ(2).

Author information:

(1)Inria/Université Bordeaux/CNRS joint project-team MAGNOME, 351, cours de la

Libération, Talence, F-33405, France. anna.zhukova@inria.fr. (2)Inria/Université

Bordeaux/CNRS joint project-team MAGNOME, 351, cours de la Libération, Talence,

F-33405, France. david.sherman@inria.fr.

BACKGROUND: The complexity of genome-scale metabolic models makes them quite

difficult for human users to read, since they contain thousands of reactions that

must be included for accurate computer simulation. Interestingly, hidden

similarities between groups of reactions can be discovered, and generalized to

reveal higher-level patterns.

RESULTS: The web-based navigation system Mimoza allows a human expert to explore

metabolic network models in a semantically zoomable manner: The most general view

represents the compartments of the model; the next view shows the generalized

versions of reactions and metabolites in each compartment; and the most detailed

view represents the initial network with the generalization-based layout (where

similar metabolites and reactions are placed next to each other). It allows a

human expert to grasp the general structure of the network and analyze it in a

top-down manner

CONCLUSIONS: Mimoza can be installed standalone, or used on-line at

http://mimoza.bordeaux.inria.fr/ , or installed in a Galaxy server for use in

workflows. Mimoza views can be embedded in web pages, or downloaded as COMBINE

archives.

DOI: 10.1186/s12918-015-0151-5

PMCID: PMC4345040

PMID: 25889977 [Indexed for MEDLINE]

624. ACS Chem Biol. 2015 Feb 20;10(2):452-9. doi: 10.1021/cb5006382. Epub 2014 Nov 5.

Unified and isomer-specific NMR metabolomics database for the accurate analysis

of (13)C-(1)H HSQC spectra.

Bingol K(1), Li DW, Bruschweiler-Li L, Cabrera OA, Megraw T, Zhang F,

Brüschweiler R.

Author information:

(1)Department of Chemistry and Biochemistry, ‡Campus Chemical Instrument Center,

The Ohio State University , Columbus, Ohio 43210, United States.

A new metabolomics database and query algorithm for the analysis of (13)C-(1)H

HSQC spectra is introduced, which unifies NMR spectroscopic information on 555

metabolites from both the Biological Magnetic Resonance Data Bank (BMRB) and

Human Metabolome Database (HMDB). The new database, termed Complex Mixture

Analysis by NMR (COLMAR) (13)C-(1)H HSQC database, can be queried via an

interactive, easy to use web interface at

http://spin.ccic.ohio-state.edu/index.php/hsqc/index . Our new HSQC database

separately treats slowly exchanging isomers that belong to the same metabolite,

which permits improved query in cases where lowly populated isomers are below the

HSQC detection limit. The performance of our new database and query web server

compares favorably with the one of existing web servers, especially for spectra

of samples of high complexity, including metabolite mixtures from the model

organisms Drosophila melanogaster and Escherichia coli. For such samples, our web

server has on average a 37% higher accuracy (true positive rate) and a 82% lower

false positive rate, which makes it a useful tool for the rapid and accurate

identification of metabolites from (13)C-(1)H HSQC spectra at natural abundance.

This information can be combined and validated with NMR data from 2D TOCSY-type

spectra that provide connectivity information not present in HSQC spectra.

DOI: 10.1021/cb5006382

PMCID: PMC4340359

PMID: 25333826 [Indexed for MEDLINE]

625. Bioinformatics. 2015 Feb 15;31(4):590-2. doi: 10.1093/bioinformatics/btu681. Epub

2014 Oct 17.

mimicMe: a web server for prediction and analysis of host-like proteins in

microbial pathogens.

Petrenko P(1), Doxey AC(1).

Author information:

(1)Department of Biology, University of Waterloo, Waterloo, Ontario, Canada.

SUMMARY: mimicMe is a web server for prediction and analysis of host-like

proteins (mimics) encoded by microbial pathogens. Users select a host species and

any set of pathogen and control proteomes (bacterial, fungal, protozoan or viral)

and mimicMe reports host-like proteins that are unique to or enriched among

pathogens. Additional server features include visualization of structural

similarities between pathogen and host proteins as well as function-enrichment

analysis.

AVAILABILITY AND IMPLEMENTATION: mimicMe is available at

http://mimicme.uwaterloo.ca

CONTACT: acdoxey@uwaterloo.ca.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu681

PMID: 25399027 [Indexed for MEDLINE]

626. Bioinformatics. 2015 Feb 15;31(4):612-3. doi: 10.1093/bioinformatics/btu688. Epub

2014 Oct 17.

IntSide: a web server for the chemical and biological examination of drug side

effects.

Juan-Blanco T(1), Duran-Frigola M(1), Aloy P(2).

Author information:

(1)Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in

Biomedicine (IRB Barcelona) and Institució Catalana de Recerca i Estudis Avançats

(ICREA), Barcelona, Spain. (2)Joint IRB-BSC-CRG Program in Computational Biology,

Institute for Research in Biomedicine (IRB Barcelona) and Institució Catalana de

Recerca i Estudis Avançats (ICREA), Barcelona, Spain Joint IRB-BSC-CRG Program in

Computational Biology, Institute for Research in Biomedicine (IRB Barcelona) and

Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

SUMMARY: Drug side effects are one of the main health threats worldwide, and an

important obstacle in drug development. Understanding how adverse reactions occur

requires knowledge on drug mechanisms at the molecular level. Despite recent

advances, the need for tools and methods that facilitate side effect anticipation

still remains. Here, we present IntSide, a web server that integrates chemical

and biological information to elucidate the molecular mechanisms underlying drug

side effects. IntSide currently catalogs 1175 side effects caused by 996 drugs,

associated with drug features divided into eight categories, belonging to either

biology or chemistry. On the biological side, IntSide reports drug targets and

off-targets, pathways, molecular functions and biological processes. From a

chemical viewpoint, it includes molecular fingerprints, scaffolds and chemical

entities. Finally, we also integrate additional biological data, such as protein

interactions and disease-related genes, to facilitate mechanistic

interpretations.

AVAILABILITY AND IMPLEMENTATION: Our data and web resource are available online

(http://intside.irbbarcelona.org/).

CONTACT: patrick.aloy@irbbarcelona.org

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu688

PMID: 25380960 [Indexed for MEDLINE]

627. Bioinformatics. 2015 Feb 15;31(4):523-31. doi: 10.1093/bioinformatics/btu673.

Epub 2014 Oct 15.

FuncPatch: a web server for the fast Bayesian inference of conserved functional

patches in protein 3D structures.

Huang YF(1), Golding GB(1).

Author information:

(1)Department of Biology, McMaster University, Hamilton, ON L8S4K1, Canada.

MOTIVATION: A number of statistical phylogenetic methods have been developed to

infer conserved functional sites or regions in proteins. Many methods, e.g.

Rate4Site, apply the standard phylogenetic models to infer site-specific

substitution rates and totally ignore the spatial correlation of substitution

rates in protein tertiary structures, which may reduce their power to identify

conserved functional patches in protein tertiary structures when the sequences

used in the analysis are highly similar. The 3D sliding window method has been

proposed to infer conserved functional patches in protein tertiary structures,

but the window size, which reflects the strength of the spatial correlation, must

be predefined and is not inferred from data. We recently developed GP4Rate to

solve these problems under the Bayesian framework. Unfortunately, GP4Rate is

computationally slow. Here, we present an intuitive web server, FuncPatch, to

perform a fast approximate Bayesian inference of conserved functional patches in

protein tertiary structures.

RESULTS: Both simulations and four case studies based on empirical data suggest

that FuncPatch is a good approximation to GP4Rate. However, FuncPatch is orders

of magnitudes faster than GP4Rate. In addition, simulations suggest that

FuncPatch is potentially a useful tool complementary to Rate4Site, but the 3D

sliding window method is less powerful than FuncPatch and Rate4Site. The

functional patches predicted by FuncPatch in the four case studies are supported

by experimental evidence, which corroborates the usefulness of FuncPatch.

AVAILABILITY AND IMPLEMENTATION: The software FuncPatch is freely available at

the web site, http://info.mcmaster.ca/yifei/FuncPatch

CONTACT: golding@mcmaster.ca

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu673

PMID: 25322839 [Indexed for MEDLINE]

628. PLoS One. 2015 Feb 13;10(2):e0117804. doi: 10.1371/journal.pone.0117804.

eCollection 2015.

An ensemble method with hybrid features to identify extracellular matrix

proteins.

Yang R(1), Zhang C(2), Gao R(1), Zhang L(1).

Author information:

(1)School of Control Science and Engineering, Shandong University, Jinan, China.

(2)School of Control Science and Engineering, Shandong University, Jinan, China;

School of Mechanical, Electrical and Information Engineering, Shandong University

at Weihai, China.

The extracellular matrix (ECM) is a dynamic composite of secreted proteins that

play important roles in numerous biological processes such as tissue

morphogenesis, differentiation and homeostasis. Furthermore, various diseases are

caused by the dysfunction of ECM proteins. Therefore, identifying these important

ECM proteins may assist in understanding related biological processes and drug

development. In view of the serious imbalance in the training dataset, a Random

Forest-based ensemble method with hybrid features is developed in this paper to

identify ECM proteins. Hybrid features are employed by incorporating sequence

composition, physicochemical properties, evolutionary and structural information.

The Information Gain Ratio and Incremental Feature Selection (IGR-IFS) methods

are adopted to select the optimal features. Finally, the resulting predictor

termed IECMP (Identify ECM Proteins) achieves an balanced accuracy of 86.4% using

the 10-fold cross-validation on the training dataset, which is much higher than

results obtained by other methods (ECMPRED: 71.0%, ECMPP: 77.8%). Moreover, when

tested on a common independent dataset, our method also achieves significantly

improved performance over ECMPP and ECMPRED. These results indicate that IECMP is

an effective method for ECM protein prediction, which has a more balanced

prediction capability for positive and negative samples. It is anticipated that

the proposed method will provide significant information to fully decipher the

molecular mechanisms of ECM-related biological processes and discover candidate

drug targets. For public access, we develop a user-friendly web server for ECM

protein identification that is freely accessible at http://iecmp.weka.cc.

DOI: 10.1371/journal.pone.0117804

PMCID: PMC4334504

PMID: 25680094 [Indexed for MEDLINE]

629. Database (Oxford). 2015 Feb 2;2015. pii: bav001. doi: 10.1093/database/bav001.

Print 2015.

EcoliNet: a database of cofunctional gene network for Escherichia coli.

Kim H(1), Shim JE(1), Shin J(1), Lee I(2).

Author information:

(1)Department of Biotechnology, College of Life Science and Biotechnology, Yonsei

University, Seoul, Korea. (2)Department of Biotechnology, College of Life Science

and Biotechnology, Yonsei University, Seoul, Korea insuklee@yonsei.ac.kr.

During the past several decades, Escherichia coli has been a treasure chest for

molecular biology. The molecular mechanisms of many fundamental cellular

processes have been discovered through research on this bacterium. Although much

basic research now focuses on more complex model organisms, E. coli still remains

important in metabolic engineering and synthetic biology. Despite its long

history as a subject of molecular investigation, more than one-third of the E.

coli genome has no pathway annotation supported by either experimental evidence

or manual curation. Recently, a network-assisted genetics approach to the

efficient identification of novel gene functions has increased in popularity. To

accelerate the speed of pathway annotation for the remaining uncharacterized part

of the E. coli genome, we have constructed a database of cofunctional gene

network with near-complete genome coverage of the organism, dubbed EcoliNet. We

find that EcoliNet is highly predictive for diverse bacterial phenotypes,

including antibiotic response, indicating that it will be useful in prioritizing

novel candidate genes for a wide spectrum of bacterial phenotypes. We have

implemented a web server where biologists can easily run network algorithms over

EcoliNet to predict novel genes involved in a pathway or novel functions for a

gene. All integrated cofunctional associations can be downloaded, enabling

orthology-based reconstruction of gene networks for other bacterial species as

well. Database URL: http://www.inetbio.org/ecolinet.

© The Author(s) 2015. Published by Oxford University Press.

DOI: 10.1093/database/bav001

PMCID: PMC4314589

PMID: 25650278 [Indexed for MEDLINE]

630. Bioinformatics. 2015 Feb 1;31(3):418-20. doi: 10.1093/bioinformatics/btu655. Epub

2014 Oct 13.

SNPsnap: a Web-based tool for identification and annotation of matched SNPs.

Pers TH(1), Timshel P(1), Hirschhorn JN(1).

Author information:

(1)Division of Endocrinology and Center for Basic and Translational Obesity

Research, Boston Children's Hospital, Boston, MA 02115, Medical and Population

Genetics Program, Broad Institute of MIT and Harvard, Cambridge, MA 2142, USA,

Department of Systems Biology, Center for Biological Sequence Analysis, Technical

University of Denmark, 2800 Lyngby, Denmark and Department of Genetics, Harvard

Medical School, Boston, MA 02115, USA Division of Endocrinology and Center for

Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA

02115, Medical and Population Genetics Program, Broad Institute of MIT and

Harvard, Cambridge, MA 2142, USA, Department of Systems Biology, Center for

Biological Sequence Analysis, Technical University of Denmark, 2800 Lyngby,

Denmark and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Division of Endocrinology and Center for Basic and Translational Obesity

Research, Boston Children's Hospital, Boston, MA 02115, Medical and Population

Genetics Program, Broad Institute of MIT and Harvard, Cambridge, MA 2142, USA,

Department of Systems Biology, Center for Biological Sequence Analysis, Technical

University of Denmark, 2800 Lyngby, Denmark and Department of Genetics, Harvard

Medical School, Boston, MA 02115, USA.

SUMMARY: An important computational step following genome-wide association

studies (GWAS) is to assess whether disease or trait-associated single-nucleotide

polymorphisms (SNPs) enrich for particular biological annotations. SNP-based

enrichment analysis needs to account for biases such as co-localization of GWAS

signals to gene-dense and high linkage disequilibrium (LD) regions, and

correlations of gene size, location and function. The SNPsnap Web server enables

SNP-based enrichment analysis by providing matched sets of SNPs that can be used

to calibrate background expectations. Specifically, SNPsnap efficiently

identifies sets of randomly drawn SNPs that are matched to a set of query SNPs

based on allele frequency, number of SNPs in LD, distance to nearest gene and

gene density.

AVAILABILITY AND IMPLEMENTATION: SNPsnap server is available at

http://www.broadinstitute.org/mpg/snpsnap/.

CONTACT: joelh@broadinstitute.org

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu655

PMCID: PMC4308663

PMID: 25316677 [Indexed for MEDLINE]

631. Bioinformatics. 2015 Feb 1;31(3):382-9. doi: 10.1093/bioinformatics/btu663. Epub

2014 Oct 9.

PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation

events.

Patrick R(1), Lê Cao KA(2), Kobe B(2), Bodén M(3).

Author information:

(1)School of Chemistry and Molecular Biosciences and Queensland Facility for

Advanced Bioinformatics, The University of Queensland, St Lucia 4072,

Translational Research Institute, The University of Queensland Diamantina

Institute, Brisbane, St Lucia 4102, Institute for Molecular Bioscience and

Australian Infectious Diseases Research Centre, The University of Queensland, St

Lucia, 4072, Australia. (2)School of Chemistry and Molecular Biosciences and

Queensland Facility for Advanced Bioinformatics, The University of Queensland, St

Lucia 4072, Translational Research Institute, The University of Queensland

Diamantina Institute, Brisbane, St Lucia 4102, Institute for Molecular Bioscience

and Australian Infectious Diseases Research Centre, The University of Queensland,

St Lucia, 4072, Australia School of Chemistry and Molecular Biosciences and

Queensland Facility for Advanced Bioinformatics, The University of Queensland, St

Lucia 4072, Translational Research Institute, The University of Queensland

Diamantina Institute, Brisbane, St Lucia 4102, Institute for Molecular Bioscience

and Australian Infectious Diseases Research Centre, The University of Queensland,

St Lucia, 4072, Australia School of Chemistry and Molecular Biosciences and

Queensland Facility for Advanced Bioinformatics, The University of Queensland, St

Lucia 4072, Translational Research Institute, The University of Queensland

Diamantina Institute, Brisbane, St Lucia 4102, Institute for Molecular Bioscience

and Australian Infectious Diseases Research Centre, The University of Queensland,

St Lucia, 4072, Australia. (3)School of Chemistry and Molecular Biosciences and

Queensland Facility for Advanced Bioinformatics, The University of Queensland, St

Lucia 4072, Translational Research Institute, The University of Queensland

Diamantina Institute, Brisbane, St Lucia 4102, Institute for Molecular Bioscience

and Australian Infectious Diseases Research Centre, The University of Queensland,

St Lucia, 4072, Australia School of Chemistry and Molecular Biosciences and

Queensland Facility for Advanced Bioinformatics, The University of Queensland, St

Lucia 4072, Translational Research Institute, The University of Queensland

Diamantina Institute, Brisbane, St Lucia 4102, Institute for Molecular Bioscience

and Australian Infectious Diseases Research Centre, The University of Queensland,

St Lucia, 4072, Australia.

MOTIVATION: The determinants of kinase-substrate phosphorylation can be found

both in the substrate sequence and the surrounding cellular context. Cell cycle

progression, interactions with mediating proteins and even prior phosphorylation

events are necessary for kinases to maintain substrate specificity. While much

work has focussed on the use of sequence-based methods to predict phosphorylation

sites, there has been very little work invested into the application of systems

biology to understand phosphorylation. Lack of specificity in many kinase

substrate binding motifs means that sequence methods for predicting kinase

binding sites are susceptible to high false-positive rates.

RESULTS: We present here a model that takes into account protein-protein

interaction information, and protein abundance data across the cell cycle to

predict kinase substrates for 59 human kinases that are representative of

important biological pathways. The model shows high accuracy for substrate

prediction (with an average AUC of 0.86) across the 59 kinases tested. When using

the model to complement sequence-based kinase-specific phosphorylation site

prediction, we found that the additional information increased prediction

performance for most comparisons made, particularly on kinases from the CMGC

family. We then used our model to identify functional overlaps between predicted

CDK2 substrates and targets from the E2F family of transcription factors. Our

results demonstrate that a model harnessing context data can account for the

short-falls in sequence information and provide a robust description of the

cellular events that regulate protein phosphorylation.

AVAILABILITY AND IMPLEMENTATION: The method is freely available online as a web

server at the website http://bioinf.scmb.uq.edu.au/phosphopick.

CONTACT: m.boden@uq.edu.au

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu663

PMID: 25304781 [Indexed for MEDLINE]

632. Bioinformatics. 2015 Feb 1;31(3):434-5. doi: 10.1093/bioinformatics/btu667. Epub

2014 Oct 9.

Tabhu: tools for antibody humanization.

Olimpieri PP(1), Marcatili P(1), Tramontano A(2).

Author information:

(1)Department of Physics, Sapienza University, Rome 00185, Italy, Center for

Biological Sequence Analysis, Department of Systems Biology, Technical University

of Denmark, Anker Engelunds Vej 1, 2800 Lyngby, Denmark and Istituto

Pasteur-Fondazione Cenci Bolognetti, Rome 00185, Italy. (2)Department of Physics,

Sapienza University, Rome 00185, Italy, Center for Biological Sequence Analysis,

Department of Systems Biology, Technical University of Denmark, Anker Engelunds

Vej 1, 2800 Lyngby, Denmark and Istituto Pasteur-Fondazione Cenci Bolognetti,

Rome 00185, Italy Department of Physics, Sapienza University, Rome 00185, Italy,

Center for Biological Sequence Analysis, Department of Systems Biology, Technical

University of Denmark, Anker Engelunds Vej 1, 2800 Lyngby, Denmark and Istituto

Pasteur-Fondazione Cenci Bolognetti, Rome 00185, Italy.

SUMMARY: Antibodies are rapidly becoming essential tools in the clinical

practice, given their ability to recognize their cognate antigens with high

specificity and affinity, and a high yield at reasonable costs in model animals.

Unfortunately, when administered to human patients, xenogeneic antibodies can

elicit unwanted and dangerous immunogenic responses. Antibody humanization

methods are designed to produce molecules with a better safety profile still

maintaining their ability to bind the antigen. This can be accomplished by

grafting the non-human regions determining the antigen specificity into a

suitable human template. Unfortunately, this procedure may results in a partial

or complete loss of affinity of the grafted molecule that can be restored by

back-mutating some of the residues of human origin to the corresponding murine

ones. This trial-and-error procedure is hard and involves expensive and

time-consuming experiments. Here we present tools for antibody humanization

(Tabhu) a web server for antibody humanization. Tabhu includes tools for human

template selection, grafting, back-mutation evaluation, antibody modelling and

structural analysis, helping the user in all the critical steps of the

humanization experiment protocol.

AVAILABILITY: http://www.biocomputing.it/tabhu

CONTACT: anna.tramontano@uniroma1.it, pierpaolo.olimpieri@uniroma1.it

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu667

PMCID: PMC4308665

PMID: 25304777 [Indexed for MEDLINE]

633. Bioinformatics. 2015 Feb 1;31(3):405-12. doi: 10.1093/bioinformatics/btu626. Epub

2014 Oct 9.

PDB-wide collection of binding data: current status of the PDBbind database.

Liu Z(1), Li Y(1), Han L(1), Li J(1), Liu J(1), Zhao Z(1), Nie W(1), Liu Y(1),

Wang R(2).

Author information:

(1)State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai

Institute of Organic Chemistry, Chinese Academy of Sciences, 345 Lingling Road,

Shanghai 200032 and State Key Laboratory of Quality Research in Chinese Medicine,

Macau Institute for Applied Research in Medicine and Health, Macau University of

Science and Technology, Macau, People's Republic of China. (2)State Key

Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of

Organic Chemistry, Chinese Academy of Sciences, 345 Lingling Road, Shanghai

200032 and State Key Laboratory of Quality Research in Chinese Medicine, Macau

Institute for Applied Research in Medicine and Health, Macau University of

Science and Technology, Macau, People's Republic of China State Key Laboratory of

Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic

Chemistry, Chinese Academy of Sciences, 345 Lingling Road, Shanghai 200032 and

State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for

Applied Research in Medicine and Health, Macau University of Science and

Technology, Macau, People's Republic of China.

MOTIVATION: Molecular recognition between biological macromolecules and organic

small molecules plays an important role in various life processes. Both

structural information and binding data of biomolecular complexes are

indispensable for depicting the underlying mechanism in such an event. The

PDBbind database was created to collect experimentally measured binding data for

the biomolecular complexes throughout the Protein Data Bank (PDB). It thus

provides the linkage between structural information and energetic properties of

biomolecular complexes, which is especially desirable for computational studies

or statistical analyses.

RESULTS: Since its first public release in 2004, the PDBbind database has been

updated on an annual basis. The latest release (version 2013) provides

experimental binding affinity data for 10,776 biomolecular complexes in PDB,

including 8302 protein-ligand complexes and 2474 other types of complexes. In

this article, we will describe the current methods used for compiling PDBbind and

the updated status of this database. We will also review some typical

applications of PDBbind published in the scientific literature.

AVAILABILITY AND IMPLEMENTATION: All contents of this database are freely

accessible at the PDBbind-CN Web server at http://www.pdbbind-cn.org/.

CONTACT: wangrx@mail.sioc.ac.cn.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu626

PMID: 25301850 [Indexed for MEDLINE]

634. Bioinformatics. 2015 Feb 1;31(3):416-7. doi: 10.1093/bioinformatics/btu645. Epub

2014 Oct 7.

DupliPHY-Web: a web server for DupliPHY and DupliPHY-ML.

Ames RM(1), Lovell SC(1).

Author information:

(1)Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester

M13 9PL, UK.

SUMMARY: Gene duplication and loss are important processes in the evolution of

gene families. Moreover, growth of families by duplication and retention is an

important mechanism by which organisms gain new functions. Therefore the ability

to infer the evolutionary histories of families is an important step in

understanding the evolution of function. We have recently developed DupliPHY, a

software tool to infer gene family histories using parsimony and maximum

likelihood. Here, we present DupliPHY-Web a web server for DupliPHY that

implements additional maximum likelihood functionality and provides users an

intuitive interface to run DupliPHY.

AVAILABILITY AND IMPLEMENTATION: DupliPHY-Web is available at

www.bioinf.manchester.ac.uk/dupliphy/

CONTACT: : ryan.ames@manchester.ac.uk

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu645

PMCID: PMC4308661

PMID: 25294920 [Indexed for MEDLINE]

635. Bioinformatics. 2015 Feb 1;31(3):423-5. doi: 10.1093/bioinformatics/btu649. Epub

2014 Oct 1.

The RNA shapes studio.

Janssen S(1), Giegerich R(1).

Author information:

(1)Practical Computer Science, Faculty of Technology, Bielefeld University,

D-33615 Bielefeld, Germany.

MOTIVATION: Abstract shape analysis, first proposed in 2004, allows one to

extract several relevant structures from the folding space of an RNA sequence,

preferable to focusing in a single structure of minimal free energy. We report

recent extensions to this approach.

RESULTS: We have rebuilt the original RNAshapes as a repository of components

that allows us to integrate several established tools for RNA structure analysis:

RNAshapes, RNAalishapes and pknotsRG, including its recent extension pKiss. As a

spin-off, we obtain heretofore unavailable functionality: e. g. with pKiss, we

can now perform abstract shape analysis for structures holding pseudoknots up to

the complexity of kissing hairpin motifs. The new tool pAliKiss can predict

kissing hairpin motifs from aligned sequences. Along with the integration, the

functionality of the tools was also extended in manifold ways.

AVAILABILITY AND IMPLEMENTATION: As before, the tool is available on the

Bielefeld Bioinformatics server at

http://bibiserv.cebitec.uni-bielefeld.de/rnashapesstudio.

CONTACT: bibi-help@cebitec.uni-bielefeld.de.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu649

PMCID: PMC4308662

PMID: 25273103 [Indexed for MEDLINE]

636. Mol Biosyst. 2015 Feb;11(2):354-60. doi: 10.1039/c4mb00569d. Epub 2014 Dec 1.

Computational characterization of parallel dimeric and trimeric coiled-coils

using effective amino acid indices.

Li C(1), Wang XF, Chen Z, Zhang Z, Song J.

Author information:

(1)Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash

University, Melbourne, VIC 3800, Australia. jiangnng.song@monash.edu.

The coiled-coil, which consists of two or more α-helices winding around each

other, is a ubiquitous and the most frequently observed protein-protein

interaction motif in nature. The coiled-coil is known for its straightforward

heptad repeat pattern and can be readily recognized based on protein primary

sequences, exhibiting a variety of oligomer states and topologies. Due to the

stable interaction formed between their α-helices, coiled-coils have been under

close scrutiny to design novel protein structures for potential applications in

the fields of material science, synthetic biology and medicine. However, their

broader application requires an in-depth and systematic analysis of the

sequence-to-structure relationship of coiled-coil folding and oligomeric

formation. In this article, we propose a new oligomerization state predictor,

termed as RFCoil, which exploits the most useful and non-redundant amino acid

indices combined with the machine learning algorithm - random forest (RF) - to

predict the oligomeric states of coiled-coil regions. Benchmarking experiments

show that RFCoil achieves an AUC (area under the ROC curve) of 0.849 on the

10-fold cross-validation test using the training dataset and 0.855 on the

independent test using the validation dataset, respectively. Performance

comparison results indicate that RFCoil outperforms the four existing predictors

LOGICOIL, PrOCoil, SCORER 2.0 and Multicoil2. Furthermore, we extract a number of

predominant rules from the trained RF model that underlie the oligomeric

formation. We also present two case studies to illustrate the applicability of

the extracted rules to the prediction of coiled-coil oligomerization state. The

RFCoil web server, source codes and datasets are freely available for academic

users at http://protein.cau.edu.cn/RFCoil/.

DOI: 10.1039/c4mb00569d

PMID: 25435395 [Indexed for MEDLINE]

637. BioData Min. 2015 Jan 31;8:3. doi: 10.1186/s13040-014-0031-3. eCollection 2015.

Prediction of protein solvent accessibility using PSO-SVR with multiple

sequence-derived features and weighted sliding window scheme.

Zhang J(1), Chen W(2), Sun P(3), Zhao X(3), Ma Z(1).

Author information:

(1)School of Computer Science and Information Technology, Northeast Normal

University, Changchun, 1300117 P.R. China. (2)School of Chemistry and Molecular

Biosciences, The University of Queensland, Brisbane, Queensland Australia.

(3)School of Computer Science and Information Technology, Northeast Normal

University, Changchun, 1300117 P.R. China ; The Engineering Laboratory for

Drug-Gene and Protein Screening, Northeast Normal University, Changchun, 130117

P.R. China.

BACKGROUND: The prediction of solvent accessibility could provide valuable clues

for analyzing protein structure and functions, such as protein 3-Dimensional

structure and B-cell epitope prediction. To fully decipher the protein-protein

interaction process, an initial but crucial step is to calculate the protein

solvent accessibility, especially when the tertiary structure of the protein is

unknown. Although some efforts have been put into the protein solvent

accessibility prediction, the performance of existing methods is far from

satisfaction.

METHODS: In order to develop the high-accuracy model, we focus on some possible

aspects concerning the prediction performance, including several sequence-derived

features, a weighted sliding window scheme and the parameters optimization of

machine learning approach. To address above issues, we take following strategies.

Firstly, we explore various features which have been observed to be associated

with the residue solvent accessibility. These discriminative features include

protein evolutionary information, predicted protein secondary structure, native

disorder, physicochemical propensities and several sequence-based structural

descriptors of residues. Secondly, the different contributions of adjacent

residues in sliding window are observed, thus a weighted sliding window scheme is

proposed to differentiate the contributions of adjacent residues on the central

residue. Thirdly, particle swarm optimization (PSO) is employed to search the

global best parameters for the proposed predictor.

RESULTS: Evaluated by 3-fold cross-validation, our method achieves the mean

absolute error (MAE) of 14.1% and the person correlation coefficient (PCC) of

0.75 for our new-compiled dataset. When compared with the state-of-the-art

prediction models in the two benchmark datasets, our method demonstrates better

performance. Experimental results demonstrate that our PSAP achieves high

performances and outperforms many existing predictors. A web server called PSAP

is built and freely available at http://59.73.198.144:8088/SolventAccessibility/.

DOI: 10.1186/s13040-014-0031-3

PMCID: PMC4608127

PMID: 26478747

638. Genome Biol Evol. 2015 Jan 29;7(3):735-49. doi: 10.1093/gbe/evv014.

Karyotype and gene order evolution from reconstructed extinct ancestors highlight

contrasts in genome plasticity of modern rosid crops.

Murat F(1), Zhang R(1), Guizard S(1), Gavranović H(2), Flores R(3), Steinbach

D(3), Quesneville H(3), Tannier E(4), Salse J(5).

Author information:

(1)INRA/UBP UMR 1095 GDEC 'Génétique, Diversité et Ecophysiologie des Céréales',

Clermont Ferrand, France. (2)Faculty of Engineering and Natural Sciences,

International University of Sarajevo, Sarajevo, Bosnia and Herzegovina. (3)INRA

'Unité de Recherche en Génomique et Informatique', Centre INRA de Versailles,

Versailles, France. (4)INRIA Rhône-Alpes, Université de Lyon 1, CNRS UMR5558,

Laboratoire Biométrie et Biologie Évolutive, Villeurbanne Cedex, France.

(5)INRA/UBP UMR 1095 GDEC 'Génétique, Diversité et Ecophysiologie des Céréales',

Clermont Ferrand, France jsalse@clermont.inra.fr.

We used nine complete genome sequences, from grape, poplar, Arabidopsis, soybean,

lotus, apple, strawberry, cacao, and papaya, to investigate the paleohistory of

rosid crops. We characterized an ancestral rosid karyotype, structured into 7/21

protochomosomes, with a minimal set of 6,250 ordered protogenes and a minimum

physical coding gene space of 50 megabases. We also proposed ancestral karyotypes

for the Caricaceae, Brassicaceae, Malvaceae, Fabaceae, Rosaceae, Salicaceae, and

Vitaceae families with 9, 8, 10, 6, 12, 9, 12, and 19 protochromosomes,

respectively. On the basis of these ancestral karyotypes and present-day species

comparisons, we proposed a two-step evolutionary scenario based on

allohexaploidization involving the newly characterized A, B, and C diploid

progenitors leading to dominant (stable) and sensitive (plastic) genomic

compartments in any modern rosid crops. Finally, a new user-friendly online tool,

"DicotSyntenyViewer" (available from

http://urgi.versailles.inra.fr/synteny-dicot), has been made available for

accurate translational genomics in rosids.

© The Author(s) 2015. Published by Oxford University Press on behalf of the

Society for Molecular Biology and Evolution.

DOI: 10.1093/gbe/evv014

PMCID: PMC5322550

PMID: 25637221 [Indexed for MEDLINE]

639. BMC Bioinformatics. 2015 Jan 28;16:20. doi: 10.1186/s12859-014-0444-5.

ERD: a fast and reliable tool for RNA design including constraints.

Esmaili-Taheri A(1), Ganjtabesh M(2,)(3).

Author information:

(1)Department of Computer Science, School of Mathematics, Statistics, and

Computer Science, College of Science, University of Tehran, Tehran, Iran.

aliesmaili@ut.ac.ir. (2)Department of Computer Science, School of Mathematics,

Statistics, and Computer Science, College of Science, University of Tehran,

Tehran, Iran. mgtabesh@ut.ac.ir. (3)Laboratoire d'Informatique (LIX), Ecole

Polytechnique, Palaiseau CEDEX, 91128, France. mgtabesh@ut.ac.ir.

BACKGROUND: The function of an RNA in cellular processes is directly related to

its structure. The free energy of RNA structure in another important key to its

function as only some structures with a specific level of free energy can take

part in cellular reactions. Therefore, to perform a specific function, a

particular RNA structure with specific level of free energy is required. For a

given RNA structure, the goal of the RNA design problem is to design an RNA

sequence that folds into the given structure. To mimic the biological features of

RNA sequences and structures, some sequence and energy constraints should be

considered in designing RNA. Although the level of free energy is important, it

is not considered in the available approaches for RNA design problem.

RESULTS: In this paper, we present a new version of our evolutionary algorithm

for RNA design problem, entitled ERD, and extend it to handle some sequence and

energy constraints. In the sequence constraints, one can restrict sequence

positions to a fixed nucleotide or to a subset of nucleotides. As for the energy

constraint, one can specify an interval for the free energy ranges of the

designed sequences. We compare our algorithm with INFO-RNA, MODENA, NUPACK, and

RNAiFold approaches for some artificial and natural RNA secondary structures and

constraints.

CONCLUSIONS: The results indicate that our algorithm outperforms the other

mentioned approaches in terms of accuracy, speedup, divergency, nucleotides

distribution, and similarity to the natural RNA sequences. Particularly, the

designed RNA sequences in our method are much more reliable and similar to the

natural counterparts. The generated sequences are more diverse and they have

closer nucleotides distribution to the natural one. The ERD tool and web server

are freely available at http://mostafa.ut.ac.ir/corna/erd-cons/ .

DOI: 10.1186/s12859-014-0444-5

PMCID: PMC4384295

PMID: 25626878 [Indexed for MEDLINE]

640. J Theor Biol. 2015 Jan 21;365:96-103. doi: 10.1016/j.jtbi.2014.10.008. Epub 2014

Oct 22.

Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition

and support vector machine.

Kumar R(1), Srivastava A(1), Kumari B(1), Kumar M(2).

Author information:

(1)Department of Biophysics, University of Delhi South Campus, New Delhi, India.

(2)Department of Biophysics, University of Delhi South Campus, New Delhi, India.

Electronic address: manish@south.du.ac.in.

β-Lactam class of antibiotics is used as major therapeutic agent against a number

of pathogenic microbes. The widespread and indiscriminate use of antibiotics to

treat bacterial infection has prompted evolution of several evading mechanisms

from the lethal effect of antibiotics. β-Lactamases are endogenously produced

enzyme that makes bacteria resistant against β-lactam antibiotics by cleaving the

β-lactam ring. On the basis of primary structures, β-lactamase family of enzymes

is divided into four classes namely A, B, C and D. Class B are metallo-enzymes

while A, C and D does not need any metal in the enzyme catalysis. In the present

study we developed a SVM based two level β-lactamases protein prediction method,

which differentiate β-lactamases from non-β-lactamases at first level and then

classify predicted β-lactamases into different classes at second level. We

evaluated performance of different input vectors namely simple amino acid

composition, Type-1 and Type-2 Chou's pseudo amino acid compositions. Comparative

performances indicated that SVM model trained on Type-1 pseudo amino acid

composition has the best performance. At first level we were able to classify

β-lactamases from non-β-lactamases with 90.63% accuracy. At second level we found

maximum accuracy of 61.82%, 89.09%, 70.91% and 70.91% of class A, class B, class

C and class D, respectively. A web-server as well as standalone, PredLactamase,

is also developed to make the method available to the scientific community, which

can be accessed at http://14.139.227.92/mkumar/predlactamase.

Copyright © 2014 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2014.10.008

PMID: 25454009 [Indexed for MEDLINE]

641. Bioinformatics. 2015 Jan 15;31(2):265-7. doi: 10.1093/bioinformatics/btu614. Epub

2014 Sep 29.

BioVLAB-MMIA-NGS: microRNA-mRNA integrated analysis using high-throughput

sequencing data.

Chae H(1), Rhee S(1), Nephew KP(1), Kim S(1).

Author information:

(1)Department of Computer Science, School of Informatics and Computing, Indiana

University Bloomington, IN 47404, USA, School of Computer Science and

Engineering, Seoul National University, Seoul, Korea and Indiana University

School of Medicine, Indianapolis, IN 46202, USA.

MOTIVATION: It is now well established that microRNAs (miRNAs) play a critical

role in regulating gene expression in a sequence-specific manner, and genome-wide

efforts are underway to predict known and novel miRNA targets. However, the

integrated miRNA-mRNA analysis remains a major computational challenge, requiring

powerful informatics systems and bioinformatics expertise.

RESULTS: The objective of this study was to modify our widely recognized Web

server for the integrated mRNA-miRNA analysis (MMIA) and its subsequent

deployment on the Amazon cloud (BioVLAB-MMIA) to be compatible with

high-throughput platforms, including next-generation sequencing (NGS) data (e.g.

RNA-seq). We developed a new version called the BioVLAB-MMIA-NGS, deployed on

both Amazon cloud and on a high-performance publicly available server called

MAHA. By using NGS data and integrating various bioinformatics tools and

databases, BioVLAB-MMIA-NGS offers several advantages. First, sequencing data is

more accurate than array-based methods for determining miRNA expression levels.

Second, potential novel miRNAs can be detected by using various computational

methods for characterizing miRNAs. Third, because miRNA-mediated gene regulation

is due to hybridization of an miRNA to its target mRNA, sequencing data can be

used to identify many-to-many relationship between miRNAs and target genes with

high accuracy.

AVAILABILITY AND IMPLEMENTATION: http://epigenomics.snu.ac.kr/biovlab\_mmia\_ngs/.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu614

PMID: 25270639 [Indexed for MEDLINE]

642. Bioinformatics. 2015 Jan 15;31(2):282-3. doi: 10.1093/bioinformatics/btu616. Epub

2014 Sep 26.

Shiny-phyloseq: Web application for interactive microbiome analysis with

provenance tracking.

McMurdie PJ(1), Holmes S(1).

Author information:

(1)Department of Statistics, Stanford University, Stanford, CA 94305, USA.

We have created a Shiny-based Web application, called Shiny-phyloseq, for dynamic

interaction with microbiome data that runs on any modern Web browser and requires

no programming, increasing the accessibility and decreasing the entrance

requirement to using phyloseq and related R tools. Along with a data- and

context-aware dynamic interface for exploring the effects of parameter and method

choices, Shiny-phyloseq also records the complete user input and subsequent

graphical results of a user's session, allowing the user to archive, share and

reproduce the sequence of steps that created their result-without writing any new

code themselves.AVAILABILITY AND IMPLEMENTATION: Shiny-phyloseq is implemented

entirely in the R language. It can be hosted/launched by any system with R

installed, including Windows, Mac OS and most Linux distributions. Information

technology administrators can also host Shiny--phyloseq from a remote server, in

which case users need only have a Web browser installed. Shiny-phyloseq is

provided free of charge under a GPL-3 open-source license through GitHub at

http://joey711.github.io/shiny-phyloseq/.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu616

PMCID: PMC4287943

PMID: 25262154 [Indexed for MEDLINE]

643. Bioinformatics. 2015 Jan 15;31(2):194-200. doi: 10.1093/bioinformatics/btu598.

Epub 2014 Sep 17.

Proteomic analysis and prediction of human phosphorylation sites in subcellular

level reveal subcellular specificity.

Chen X(1), Shi SP(2), Suo SB(1), Xu HD(1), Qiu JD(2).

Author information:

(1)Department of Chemistry, Nanchang University, Nanchang 330031, Department of

Mathematics, Nanchang University, Nanchang 330031 and Department of Materials and

Chemical Engineering, Pingxiang College, Pingxiang 337055, P.R. China.

(2)Department of Chemistry, Nanchang University, Nanchang 330031, Department of

Mathematics, Nanchang University, Nanchang 330031 and Department of Materials and

Chemical Engineering, Pingxiang College, Pingxiang 337055, P.R. China Department

of Chemistry, Nanchang University, Nanchang 330031, Department of Mathematics,

Nanchang University, Nanchang 330031 and Department of Materials and Chemical

Engineering, Pingxiang College, Pingxiang 337055, P.R. China.

MOTIVATION: Protein phosphorylation is the most common post-translational

modification (PTM) regulating major cellular processes through highly dynamic and

complex signaling pathways. Large-scale comparative phosphoproteomic studies have

frequently been done on whole cells or organs by conventional bottom-up mass

spectrometry approaches, i.e at the phosphopeptide level. Using this approach,

there is no way to know from where the phosphopeptide signal originated. Also, as

a consequence of the scale of these studies, important information on the

localization of phosphorylation sites in subcellular compartments (SCs) is not

surveyed.

RESULTS: Here, we present a first account of the emerging field of subcellular

phosphoproteomics where a support vector machine (SVM) approach was combined with

a novel algorithm of discrete wavelet transform (DWT) to facilitate the

identification of compartment-specific phosphorylation sites and to unravel the

intricate regulation of protein phosphorylation. Our data reveal that the

subcellular phosphorylation distribution is compartment type dependent and that

the phosphorylation displays site-specific sequence motifs that diverge between

SCs.

AVAILABILITY AND IMPLEMENTATION: The method and database both are available as a

web server at: http://bioinfo.ncu.edu.cn/SubPhos.aspx.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu598

PMID: 25236462 [Indexed for MEDLINE]

644. PeerJ. 2015 Jan 13;3:e725. doi: 10.7717/peerj.725. eCollection 2015.

DrugOn: a fully integrated pharmacophore modeling and structure optimization

toolkit.

Vlachakis D(1), Fakourelis P(1), Megalooikonomou V(2), Makris C(2), Kossida S(3).

Author information:

(1)Bioinformatics & Medical Informatics Team, Biomedical Research Foundation,

Academy of Athens , Athens , Greece ; Computer Engineering and Informatics

Department, University of Patras , Patras , Greece. (2)Computer Engineering and

Informatics Department, University of Patras , Patras , Greece. (3)Bioinformatics

& Medical Informatics Team, Biomedical Research Foundation, Academy of Athens ,

Athens , Greece ; IMGT, Laboratoire d'ImmunoGénétique Moléculaire, Institut de

Génétique Humaine , Montpellier , France.

During the past few years, pharmacophore modeling has become one of the key

components in computer-aided drug design and in modern drug discovery. DrugOn is

a fully interactive pipeline designed to exploit the advantages of modern

programming and overcome the command line barrier with two friendly environments

for the user (either novice or experienced in the field of Computer Aided Drug

Design) to perform pharmacophore modeling through an efficient combination of the

PharmACOphore, Gromacs, Ligbuilder and PDB2PQR suites. Our platform features a

novel workflow that guides the user through each logical step of the iterative 3D

structural optimization setup and drug design process. For the pharmacophore

modeling we are focusing on either the characteristics of the receptor or the

full molecular system, including a set of selected ligands. DrugOn can be freely

downloaded from our dedicated server system at

www.bioacademy.gr/bioinformatics/drugon/.

DOI: 10.7717/peerj.725

PMCID: PMC4304849

PMID: 25648563

645. Bioinformatics. 2015 Jan 1;31(1):143-5. doi: 10.1093/bioinformatics/btu613. Epub

2014 Sep 17.

The Ensembl REST API: Ensembl Data for Any Language.

Yates A(1), Beal K(1), Keenan S(1), McLaren W(1), Pignatelli M(1), Ritchie GR(2),

Ruffier M(1), Taylor K(1), Vullo A(1), Flicek P(2).

Author information:

(1)European Molecular Biology Laboratory, European Bioinformatics Institute,

Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD and Wellcome Trust

Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

(2)European Molecular Biology Laboratory, European Bioinformatics Institute,

Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD and Wellcome Trust

Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

European Molecular Biology Laboratory, European Bioinformatics Institute,

Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD and Wellcome Trust

Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

MOTIVATION: We present a Web service to access Ensembl data using

Representational State Transfer (REST). The Ensembl REST server enables the easy

retrieval of a wide range of Ensembl data by most programming languages, using

standard formats such as JSON and FASTA while minimizing client work. We also

introduce bindings to the popular Ensembl Variant Effect Predictor tool

permitting large-scale programmatic variant analysis independent of any specific

programming language.

AVAILABILITY AND IMPLEMENTATION: The Ensembl REST API can be accessed at

http://rest.ensembl.org and source code is freely available under an Apache 2.0

license from http://github.com/Ensembl/ensembl-rest.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu613

PMCID: PMC4271150

PMID: 25236461 [Indexed for MEDLINE]

646. Bioinformatics. 2015 Jan 1;31(1):119-20. doi: 10.1093/bioinformatics/btu602. Epub

2014 Sep 16.

PseKNC-General: a cross-platform package for generating various modes of pseudo

nucleotide compositions.

Chen W(1), Zhang X(2), Brooker J(2), Lin H(3), Zhang L(2), Chou KC(1).

Author information:

(1)Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, Hebei United University, Tangshan 063009, China,

Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, School of

Life Science and Technology, Bioinformatics and Computer-Aided Drug Discovery,

Gordon Life Science Institute, Boston, MA 02478, Department of Electrical and

Computer Engineering, University of Virginia, Charlottesville, VA 22904,

Department of Computer Science, Vassar College, Poughkeepsie, NY 12604, USA,

Excellence in Genomic Medicine Research, Key Laboratory for Neuro-Information of

Ministry of Education, Center of Bioinformatics, School of Life Science and

Technology, University of Electronic Science and Technology of China, Chengdu

610054, China and Excellence in Genomic Medicine Research, Center of Excellence

in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589,

Saudi Arabia Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, Hebei United University, Tangshan 063009, China,

Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, School of

Life Science and Technology, Bioinformatics and Computer-Aided Drug Discovery,

Gordon Life Science Institute, Boston, MA 02478, Department of Electrical and

Computer Engineering, University of Virginia, Charlottesville, VA 22904,

Department of Computer Science, Vassar College, Poughkeepsie, NY 12604, USA,

Excellence in Genomic Medicine Research, Key Laboratory for Neuro-Information of

Ministry of Education, Center of Bioinformatics, School of Life Science and

Technology, University of Electronic Science and Technology of China, Chengdu

610054, China and Excellence in Genomic Medicine Research, Center of Excellence

in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589,

Saudi Arabia Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, Hebei United University, Tangshan 063009, Chin

(2)Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, Hebei United University, Tangshan 063009, China,

Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, School of

Life Science and Technology, Bioinformatics and Computer-Aided Drug Discovery,

Gordon Life Science Institute, Boston, MA 02478, Department of Electrical and

Computer Engineering, University of Virginia, Charlottesville, VA 22904,

Department of Computer Science, Vassar College, Poughkeepsie, NY 12604, USA,

Excellence in Genomic Medicine Research, Key Laboratory for Neuro-Information of

Ministry of Education, Center of Bioinformatics, School of Life Science and

Technology, University of Electronic Science and Technology of China, Chengdu

610054, China and Excellence in Genomic Medicine Research, Center of Excellence

in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589,

Saudi Arabia. (3)Department of Physics, School of Sciences, Center for Genomics

and Computational Biology, Hebei United University, Tangshan 063009, China,

Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, School of

Life Science and Technology, Bioinformatics and Computer-Aided Drug Discovery,

Gordon Life Science Institute, Boston, MA 02478, Department of Electrical and

Computer Engineering, University of Virginia, Charlottesville, VA 22904,

Department of Computer Science, Vassar College, Poughkeepsie, NY 12604, USA,

Excellence in Genomic Medicine Research, Key Laboratory for Neuro-Information of

Ministry of Education, Center of Bioinformatics, School of Life Science and

Technology, University of Electronic Science and Technology of China, Chengdu

610054, China and Excellence in Genomic Medicine Research, Center of Excellence

in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589,

Saudi Arabia Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, Hebei United University, Tangshan 063009, China,

Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, School of

Life Science and Technology, Bioinformatics and Computer-Aided Drug Discovery,

Gordon Life Science Institute, Boston, MA 02478, Department of Electrical and

Computer Engineering, University of Virginia, Charlottesville, VA 22904,

Department of Computer Science, Vassar College, Poughkeepsie, NY 12604, USA,

Excellence in Genomic Medicine Research, Key Laboratory for Neuro-Information of

Ministry of Education, Center of Bioinformatics, School of Life Science and

Technology, University of Electronic Science and Technology of China, Chengdu

610054, China and Excellence in Genomic Medicine Research, Center of Excellence

in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589,

Saudi Arabia.

SUMMARY: The avalanche of genomic sequences generated in the post-genomic age

requires efficient computational methods for rapidly and accurately identifying

biological features from sequence information. Towards this goal, we developed a

freely available and open-source package, called PseKNC-General (the general form

of pseudo k-tuple nucleotide composition), that allows for fast and accurate

computation of all the widely used nucleotide structural and physicochemical

properties of both DNA and RNA sequences. PseKNC-General can generate several

modes of pseudo nucleotide compositions, including conventional k-tuple

nucleotide compositions, Moreau-Broto autocorrelation coefficient, Moran

autocorrelation coefficient, Geary autocorrelation coefficient, Type I PseKNC and

Type II PseKNC. In every mode, >100 physicochemical properties are available for

choosing. Moreover, it is flexible enough to allow the users to calculate PseKNC

with user-defined properties. The package can be run on Linux, Mac and Windows

systems and also provides a graphical user interface.

AVAILABILITY AND IMPLEMENTATION: The package is freely available at:

http://lin.uestc.edu.cn/server/pseknc.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu602

PMID: 25231908 [Indexed for MEDLINE]

647. Bioinformatics. 2015 Jan 1;31(1):131-3. doi: 10.1093/bioinformatics/btu599. Epub

2014 Sep 5.

BalestraWeb: efficient online evaluation of drug-target interactions.

Cobanoglu MC(1), Oltvai ZN(1), Taylor DL(1), Bahar I(2).

Author information:

(1)Department of Computational and Systems Biology, University of Pittsburgh,

School of Medicine, Pittsburgh, PA, The University of Pittsburgh Drug Discovery

Institute, Pittsburgh, PA 15260 and Department of Pathology, University of

Pittsburgh, School of Medicine, Pittsburgh, PA 15213, USA Department of

Computational and Systems Biology, University of Pittsburgh, School of Medicine,

Pittsburgh, PA, The University of Pittsburgh Drug Discovery Institute,

Pittsburgh, PA 15260 and Department of Pathology, University of Pittsburgh,

School of Medicine, Pittsburgh, PA 15213, USA. (2)Department of Computational and

Systems Biology, University of Pittsburgh, School of Medicine, Pittsburgh, PA,

The University of Pittsburgh Drug Discovery Institute, Pittsburgh, PA 15260 and

Department of Pathology, University of Pittsburgh, School of Medicine,

Pittsburgh, PA 15213, USA.

SUMMARY: BalestraWeb is an online server that allows users to instantly make

predictions about the potential occurrence of interactions between any given

drug-target pair, or predict the most likely interaction partners of any drug or

target listed in the DrugBank. It also permits users to identify most similar

drugs or most similar targets based on their interaction patterns. Outputs help

to develop hypotheses about drug repurposing as well as potential side effects.

AVAILABILITY AND IMPLEMENTATION: BalestraWeb is accessible at

http://balestra.csb.pitt.edu/. The tool is built using a probabilistic matrix

factorization method and DrugBank v3, and the latent variable models are trained

using the GraphLab collaborative filtering toolkit. The server is implemented

using Python, Flask, NumPy and SciPy.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu599

PMCID: PMC4271144

PMID: 25192741 [Indexed for MEDLINE]

648. Bioinformatics. 2015 Jan 1;31(1):112-3. doi: 10.1093/bioinformatics/btu547. Epub

2014 Sep 4.

mod\_bio: Apache modules for Next-Generation sequencing data.

Lindenbaum P(1), Redon R(1).

Author information:

(1)Institut National de la Santé et de la Recherche Médicale (INSERM) Unité Mixte

de Recherche (UMR) 1087, L'Institut du Thorax, Centre National de la Recherche

Scientifique (CNRS) UMR 6291, Centre Hospitalier Universitaire (CHU) de Nantes,

L'Institut du Thorax, Service de Cardiologie, 44000 Nantes and Université de

Nantes, 44000 Nantes, France Institut National de la Santé et de la Recherche

Médicale (INSERM) Unité Mixte de Recherche (UMR) 1087, L'Institut du Thorax,

Centre National de la Recherche Scientifique (CNRS) UMR 6291, Centre Hospitalier

Universitaire (CHU) de Nantes, L'Institut du Thorax, Service de Cardiologie,

44000 Nantes and Université de Nantes, 44000 Nantes, France Institut National de

la Santé et de la Recherche Médicale (INSERM) Unité Mixte de Recherche (UMR)

1087, L'Institut du Thorax, Centre National de la Recherche Scientifique (CNRS)

UMR 6291, Centre Hospitalier Universitaire (CHU) de Nantes, L'Institut du Thorax,

Service de Cardiologie, 44000 Nantes and Université de Nantes, 44000 Nantes,

France Institut National de la Santé et de la Recherche Médicale (INSERM) Unité

Mixte de Recherche (UMR) 1087, L'Institut du Thorax, Centre National de la

Recherche Scientifique (CNRS) UMR 6291, Centre Hospitalier Universitaire (CHU) de

Nantes, L'Institut du Thorax, Service de Cardiologie, 44000 Nantes and Université

de Nantes, 44000 Nantes, France.

SUMMARY: We describe mod\_bio, a set of modules for the Apache HTTP server that

allows the users to access and query fastq, tabix, fasta and bam files through a

Web browser. Those data are made available in plain text, HTML, XML, JSON and

JSON-P. A javascript-based genome browser using the JSON-P communication

technique is provided as an example of cross-domain Web service.

AVAILABILITY AND IMPLEMENTATION: https://github.com/lindenb/mod\_bio.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu547

PMID: 25192739 [Indexed for MEDLINE]

649. Bioinformatics. 2015 Jan 1;31(1):10-6. doi: 10.1093/bioinformatics/btu595. Epub

2014 Sep 3.

BigDataScript: a scripting language for data pipelines.

Cingolani P(1), Sladek R(2), Blanchette M(2).

Author information:

(1)McGill University School of Computer Science, 3480 University Street,

Montreal, Québec H3A 0E9 and McGill University and Génome Québec Innovation

Centre, 740 Dr. Penfield Avenue, Montréal, Québec H3A 0G1, Canada McGill

University School of Computer Science, 3480 University Street, Montreal, Québec

H3A 0E9 and McGill University and Génome Québec Innovation Centre, 740 Dr.

Penfield Avenue, Montréal, Québec H3A 0G1, Canada. (2)McGill University School of

Computer Science, 3480 University Street, Montreal, Québec H3A 0E9 and McGill

University and Génome Québec Innovation Centre, 740 Dr. Penfield Avenue,

Montréal, Québec H3A 0G1, Canada.

MOTIVATION: The analysis of large biological datasets often requires complex

processing pipelines that run for a long time on large computational

infrastructures. We designed and implemented a simple script-like programming

language with a clean and minimalist syntax to develop and manage pipeline

execution and provide robustness to various types of software and hardware

failures as well as portability.

RESULTS: We introduce the BigDataScript (BDS) programming language for data

processing pipelines, which improves abstraction from hardware resources and

assists with robustness. Hardware abstraction allows BDS pipelines to run without

modification on a wide range of computer architectures, from a small laptop to

multi-core servers, server farms, clusters and clouds. BDS achieves robustness by

incorporating the concepts of absolute serialization and lazy processing, thus

allowing pipelines to recover from errors. By abstracting pipeline concepts at

programming language level, BDS simplifies implementation, execution and

management of complex bioinformatics pipelines, resulting in reduced development

and debugging cycles as well as cleaner code.

AVAILABILITY AND IMPLEMENTATION: BigDataScript is available under open-source

license at http://pcingola.github.io/BigDataScript.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu595

PMCID: PMC4271142

PMID: 25189778 [Indexed for MEDLINE]

650. Bioinformatics. 2015 Jan 1;31(1):123-5. doi: 10.1093/bioinformatics/btu594. Epub

2014 Sep 2.

CCharPPI web server: computational characterization of protein-protein

interactions from structure.

Moal IH(1), Jiménez-García B(1), Fernández-Recio J(1).

Author information:

(1)Joint BSC-IRB Research Programme in Computational Biology, Department of Life

Sciences, Barcelona Supercomputing Center, C/Jordi Girona 29, 08034 Barcelona,

Spain.

SUMMARY: The atomic structures of protein-protein interactions are central to

understanding their role in biological systems, and a wide variety of biophysical

functions and potentials have been developed for their characterization and the

construction of predictive models. These tools are scattered across a multitude

of stand-alone programs, and are often available only as model parameters

requiring reimplementation. This acts as a significant barrier to their

widespread adoption. CCharPPI integrates many of these tools into a single web

server. It calculates up to 108 parameters, including models of electrostatics,

desolvation and hydrogen bonding, as well as interface packing and

complementarity scores, empirical potentials at various resolutions, docking

potentials and composite scoring functions.

AVAILABILITY AND IMPLEMENTATION: The server does not require registration by the

user and is freely available for non-commercial academic use at

http://life.bsc.es/pid/ccharppi.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu594

PMID: 25183488 [Indexed for MEDLINE]

651. BMC Bioinformatics. 2015;16 Suppl 13:S4. doi: 10.1186/1471-2105-16-S13-S4. Epub

2015 Sep 25.

DMAP: a connectivity map database to enable identification of novel drug

repositioning candidates.

Huang H, Nguyen T, Ibrahim S, Shantharam S, Yue Z, Chen JY.

BACKGROUND: Drug repositioning is a cost-efficient and time-saving process to

drug development compared to traditional techniques. A systematic method to drug

repositioning is to identify candidate drug's gene expression profiles on target

disease models and determine how similar these profiles are to approved drugs.

Databases such as the CMAP have been developed recently to help with systematic

drug repositioning.

METHODS: To overcome the limitation of connectivity maps on data coverage, we

constructed a comprehensive in silico drug-protein connectivity map called DMAP,

which contains directed drug-to-protein effects and effect scores. The

drug-to-protein effect scores are compiled from all database entries between the

drug and protein have been previously observed and provide a confidence measure

on the quality of such drug-to-protein effects.

RESULTS: In DMAP, we have compiled the direct effects between 24,121 PubChem

Compound ID (CID), which were mapped from 289,571 chemical entities recognized

from public literature, and 5,196 reviewed Uniprot proteins. DMAP compiles a

total of 438,004 chemical-to-protein effect relationships. Compared to CMAP, DMAP

shows an increase of 221 folds in the number of chemicals and 1.92 fold in the

number of ATC codes. Furthermore, by overlapping DMAP chemicals with the approved

drugs with known indications from the TTD database and literature, we obtained

982 drugs and 622 diseases; meanwhile, we only obtained 394 drugs with known

indication from CMAP. To validate the feasibility of applying new DMAP for

systematic drug repositioning, we compared the performance of DMAP and the

well-known CMAP database on two popular computational techniques:

drug-drug-similarity-based method with leave-one-out validation and

Kolmogorov-Smirnov scoring based method. In drug-drug-similarity-based method,

the drug repositioning prediction using DMAP achieved an Area-Under-Curve (AUC)

score of 0.82, compared with that using CMAP, AUC = 0.64. For Kolmogorov-Smirnov

scoring based method, with DMAP, we were able to retrieve several drug

indications which could not be retrieved using CMAP. DMAP data can be queried

using the existing C2MAP server or downloaded freely at:

http://bio.informatics.iupui.edu/cmaps

CONCLUSIONS: Reliable measurements of how drug affect disease-related proteins

are critical to ongoing drug development in the genome medicine era. We

demonstrated that DMAP can help drug development professionals assess

drug-to-protein relationship data and improve chances of success for systematic

drug repositioning efforts.

DOI: 10.1186/1471-2105-16-S13-S4

PMCID: PMC4597058

PMID: 26423722 [Indexed for MEDLINE]

652. BMC Genomics. 2015;16 Suppl 8:S9. doi: 10.1186/1471-2164-16-S8-S9. Epub 2015 Jun

18.

Differential Evolution approach to detect recent admixture.

Kozlov K, Chebotarev D, Hassan M, Triska M, Triska P, Flegontov P, Tatarinova TV.

The genetic structure of human populations is extraordinarily complex and of

fundamental importance to studies of anthropology, evolution, and medicine. As

increasingly many individuals are of mixed origin, there is an unmet need for

tools that can infer multiple origins. Misclassification of such individuals can

lead to incorrect and costly misinterpretations of genomic data, primarily in

disease studies and drug trials. We present an advanced tool to infer ancestry

that can identify the biogeographic origins of highly mixed individuals. reAdmix

can incorporate individual's knowledge of ancestors (e.g. having some ancestors

from Turkey or a Scottish grandmother). reAdmix is an online tool available at

http://chcb.saban-chla.usc.edu/reAdmix/.

DOI: 10.1186/1471-2164-16-S8-S9

PMCID: PMC4480842

PMID: 26111206 [Indexed for MEDLINE]

653. BMC Genomics. 2015;16 Suppl 3:S6. doi: 10.1186/1471-2164-16-S3-S6. Epub 2015 Jan

29.

PNImodeler: web server for inferring protein-binding nucleotides from sequence

data.

Im J, Tuvshinjargal N, Park B, Lee W, Huang DS, Han K.

BACKGROUND: Interactions between DNA and proteins are essential to many

biological processes such as transcriptional regulation and DNA replication. With

the increased availability of structures of protein-DNA complexes, several

computational studies have been conducted to predict DNA binding sites in

proteins. However, little attempt has been made to predict protein binding sites

in DNA.

RESULTS: From an extensive analysis of protein-DNA complexes, we identified

powerful features of DNA and protein sequences which can be used in predicting

protein binding sites in DNA sequences. We developed two support vector machine

(SVM) models that predict protein binding nucleotides from DNA and/or protein

sequences. One SVM model that used DNA sequence data alone achieved a sensitivity

of 73.4%, a specificity of 64.8%, an accuracy of 68.9% and a correlation

coefficient of 0.382 with a test dataset that was not used in training. Another

SVM model that used both DNA and protein sequences achieved a sensitivity of

67.6%, a specificity of 74.3%, an accuracy of 71.4% and a correlation coefficient

of 0.418.

CONCLUSIONS: Predicting binding sites in double-stranded DNAs is a more difficult

task than predicting binding sites in single-stranded molecules. Our study showed

that protein binding sites in double-stranded DNA molecules can be predicted with

a comparable accuracy as those in single-stranded molecules. Our study also

demonstrated that using both DNA and protein sequences resulted in a better

prediction performance than using DNA sequence data alone. The SVM models and

datasets constructed in this study are available at

http://bclab.inha.ac.kr/pnimodeler.

DOI: 10.1186/1471-2164-16-S3-S6

PMCID: PMC4331809

PMID: 25708089 [Indexed for MEDLINE]

654. BMC Med Genomics. 2015;8 Suppl 4:S3. doi: 10.1186/1755-8794-8-S4-S3. Epub 2015

Dec 9.

Prediction of linear B-cell epitopes of hepatitis C virus for vaccine

development.

Huang WL, Tsai MJ, Hsu KT, Wang JR, Chen YH, Ho SY.

BACKGROUND: High genetic heterogeneity in the hepatitis C virus (HCV) is the

major challenge of the development of an effective vaccine. Existing studies for

developing HCV vaccines have mainly focused on T-cell immune response. However,

identification of linear B-cell epitopes that can stimulate B-cell response is

one of the major tasks of peptide-based vaccine development. Owing to the

variability in B-cell epitope length, the prediction of B-cell epitopes is much

more complex than that of T-cell epitopes. Furthermore, the motifs of linear

B-cell epitopes in different pathogens are quite different (e. g. HCV and

hepatitis B virus). To cope with this challenge, this work aims to propose an

HCV-customized sequence-based prediction method to identify B-cell epitopes of

HCV.

RESULTS: This work establishes an experimentally verified dataset comprising the

B-cell response of HCV dataset consisting of 774 linear B-cell epitopes and 774

non B-cell epitopes from the Immune Epitope Database. An interpretable rule

mining system of B-cell epitopes (IRMS-BE) is proposed to select informative

physicochemical properties (PCPs) and then extracts several if-then rule-based

knowledge for identifying B-cell epitopes. A web server Bcell-HCV was implemented

using an SVM with the 34 informative PCPs, which achieved a training accuracy of

79.7% and test accuracy of 70.7% better than the SVM-based methods for

identifying B-cell epitopes of HCV and the two general-purpose methods. This work

performs advanced analysis of the 34 informative properties, and the results

indicate that the most effective property is the alpha-helix structure of

epitopes, which influences the connection between host cells and the E2 proteins

of HCV. Furthermore, 12 interpretable rules are acquired from top-five PCPs and

achieve a sensitivity of 75.6% and specificity of 71.3%. Finally, a conserved

promising vaccine candidate, PDREMVLYQE, is identified for inclusion in a vaccine

against HCV.

CONCLUSIONS: This work proposes an interpretable rule mining system IRMS-BE for

extracting interpretable rules using informative physicochemical properties and a

web server Bcell-HCV for predicting linear B-cell epitopes of HCV. IRMS-BE may

also apply to predict B-cell epitopes for other viruses, which benefits the

improvement of vaccines development of these viruses without significant

modification. Bcell-HCV is useful for identifying B-cell epitopes of HCV antigen

to help vaccine development, which is available at

http://e045.life.nctu.edu.tw/BcellHCV.

DOI: 10.1186/1755-8794-8-S4-S3

PMCID: PMC4682406

PMID: 26680271 [Indexed for MEDLINE]

655. BMC Med Genomics. 2015;8 Suppl 2:S8. doi: 10.1186/1755-8794-8-S2-S8. Epub 2015

May 29.

RCARE: RNA Sequence Comparison and Annotation for RNA Editing.

Lee SY, Joung JG, Park CH, Park JH, Kim JH.

The post-transcriptional sequence modification of transcripts through RNA editing

is an important mechanism for regulating protein function and is associated with

human disease phenotypes. The identification of RNA editing or RNA-DNA difference

(RDD) sites is a fundamental step in the study of RNA editing. However, a

substantial number of false-positive RDD sites have been identified recently. A

major challenge in identifying RDD sites is to distinguish between the true RNA

editing sites and the false positives. Furthermore, determining the location of

condition-specific RDD sites and elucidating their functional roles will help

toward understanding various biological phenomena that are mediated by RNA

editing. The present study developed RNA-sequence comparison and annotation for

RNA editing (RCARE) for searching, annotating, and visualizing RDD sites using

thousands of previously known editing sites, which can be used for comparative

analyses between multiple samples. RCARE also provides evidence for improving the

reliability of identified RDD sites. RCARE is a web-based comparison, annotation,

and visualization tool, which provides rich biological annotations and useful

summary plots. The developers of previous tools that identify or annotate

RNA-editing sites seldom mention the reliability of their respective tools. In

order to address the issue, RCARE utilizes a number of scientific publications

and databases to find specific documentations respective to a particular

RNA-editing site, which generates evidence levels to convey the reliability of

RCARE. Sequence-based alignment files can be converted into VCF files using a

Python script and uploaded to the RCARE server for further analysis. RCARE is

available for free at http://www.snubi.org/software/rcare/.

DOI: 10.1186/1755-8794-8-S2-S8

PMCID: PMC4460956

PMID: 26043858 [Indexed for MEDLINE]

656. BMC Syst Biol. 2015;9 Suppl 1:S10. doi: 10.1186/1752-0509-9-S1-S10. Epub 2015 Feb

6.

Identifying DNA-binding proteins by combining support vector machine and PSSM

distance transformation.

Xu R, Zhou J, Wang H, He Y, Wang X, Liu B.

BACKGROUND: DNA-binding proteins play a pivotal role in various intra- and

extra-cellular activities ranging from DNA replication to gene expression

control. Identification of DNA-binding proteins is one of the major challenges in

the field of genome annotation. There have been several computational methods

proposed in the literature to deal with the DNA-binding protein identification.

However, most of them can't provide an invaluable knowledge base for our

understanding of DNA-protein interactions.

RESULTS: We firstly presented a new protein sequence encoding method called PSSM

Distance Transformation, and then constructed a DNA-binding protein

identification method (SVM-PSSM-DT) by combining PSSM Distance Transformation

with support vector machine (SVM). First, the PSSM profiles are generated by

using the PSI-BLAST program to search the non-redundant (NR) database. Next, the

PSSM profiles are transformed into uniform numeric representations appropriately

by distance transformation scheme. Lastly, the resulting uniform numeric

representations are inputted into a SVM classifier for prediction. Thus whether a

sequence can bind to DNA or not can be determined. In benchmark test on 525

DNA-binding and 550 non DNA-binding proteins using jackknife validation, the

present model achieved an ACC of 79.96%, MCC of 0.622 and AUC of 86.50%. This

performance is considerably better than most of the existing state-of-the-art

predictive methods. When tested on a recently constructed independent dataset

PDB186, SVM-PSSM-DT also achieved the best performance with ACC of 80.00%, MCC of

0.647 and AUC of 87.40%, and outperformed some existing state-of-the-art methods.

CONCLUSIONS: The experiment results demonstrate that PSSM Distance Transformation

is an available protein sequence encoding method and SVM-PSSM-DT is a useful tool

for identifying the DNA-binding proteins. A user-friendly web-server of

SVM-PSSM-DT was constructed, which is freely accessible to the public at the

web-site on http://bioinformatics.hitsz.edu.cn/PSSM-DT/.

DOI: 10.1186/1752-0509-9-S1-S10

PMCID: PMC4331676

PMID: 25708928 [Indexed for MEDLINE]

657. Clin Lab. 2015;61(7):749-59.

Molecular Characterization of KRAS, BRAF, and EGFR Genes in Cases with Prostatic

Adenocarcinoma; Reporting Bioinformatics Description and Recurrent Mutations.

Salmaninejad A, Ghadami S, Dizaji MZ, Golchehre Z, Estiar MA, Zamani MR,

Ebrahimzadeh-Vesal R, Nowroozi MR, Shakoori A.

BACKGROUND: Prostate cancer is one of the most common cancers which develops by

mutations and/or other genetic alterations in specific genes. Regarding the

previous studies in literature predominant mutations take place in KRAS, BRAF,

and EGFR genes in special types of cancers. In this research, we attempt to

identify the prevalence and significant role of the possible mutations in EGFR

exons 18-21, KRAS codon 12, 13, and 61, and BRAF codon 600 mutations in tumoral

tissue specimens from patients with prostatic adenocarcinoma. Furthermore, in

this research, it has been attempted to investigate the molecular characteristics

of these genes in terms of bioinformatic aspects.

METHODS: A total of 35 prostatic adenocarcinoma fresh tissue samples, enriched in

neoplastic cells, were obtained from the Cancer Institute of Iran. The presence

of mutations at codons 12, 13 and 61 of KRAS, codon 600 of BRAF and EGFR exons

18-21 were determined by direct Sanger sequencing. To evaluate the molecular

features, structure, and post-translation modification of those genes, a

bioinformatics survey was performed using the SWISS-MODEL

(http://swissmodel.expasy.org) and NetPhos 2.0

(http://www.cbs.dtu.dk/services/NetPhos/) Server. Also, using bioinformatics

software, the phylogeny tree of the mutations was drawn.

RESULTS: Mutations of codons 12 and 13 of KRAS were found in 2 of the 35

prostatic adenocarcinomas. Two cases carried homozygous mutations on exon 2 in

codon 12 (G12V) and codon 13 (G13D). Also, no mutation was detected at BRAF codon

600 and EGFR exons 18-21 in any of the samples.

CONCLUSIONS: Based on the group of patients with prostate adenocarcinoma, our

research shows that the mutations in codons 12 and 13 of KRAS are the most common

in prostate carcinomas. Noting these results and the molecular pathway of this

gene, there is a possible more perceptible role for this gene in the pathogenesis

of prostatic carcinoma. However, according to our finding, as in previous

studies, the role of BRAF and EGFR gene mutations in prostate adenocarcinoma are

less than in the KRAS gene and, therefore, we assume that these common mutations

of the KRAS gene can be used as an early determining marker for early diagnosis

of prostate adenocarcinoma. In the future, due to the complexity of etiological

parameters in prostate cancer development, the case specific tumor molecular

identification and treatment for each affected subject are urgently needed.

PMID: 26299074 [Indexed for MEDLINE]

658. Comb Chem High Throughput Screen. 2015;18(8):735-50.

Bio-AIMS Collection of Chemoinformatics Web Tools based on Molecular Graph

Information and Artificial Intelligence Models.

Munteanu CR(1), Gonzalez-Diaz H, Garcia R, Loza M, Pazos A.

Author information:

(1)Information and Communication Technologies Department, Faculty of Computer

Science, University of A Coruna, 15071 A Coruna, Spain. crm.publish@gmail.com.

The molecular information encoding into molecular descriptors is the first step

into in silico Chemoinformatics methods in Drug Design. The Machine Learning

methods are a complex solution to find prediction models for specific biological

properties of molecules. These models connect the molecular structure information

such as atom connectivity (molecular graphs) or physical-chemical properties of

an atom/group of atoms to the molecular activity (Quantitative Structure -

Activity Relationship, QSAR). Due to the complexity of the proteins, the

prediction of their activity is a complicated task and the interpretation of the

models is more difficult. The current review presents a series of 11 prediction

models for proteins, implemented as free Web tools on an Artificial Intelligence

Model Server in Biosciences, Bio-AIMS (http://bio-aims.udc.es/TargetPred.php).

Six tools predict protein activity, two models evaluate drug - protein target

interactions and the other three calculate protein - protein interactions. The

input information is based on the protein 3D structure for nine models, 1D

peptide amino acid sequence for three tools and drug SMILES formulas for two

servers. The molecular graph descriptor-based Machine Learning models could be

useful tools for in silico screening of new peptides/proteins as future drug

targets for specific treatments.

PMID: 26234511 [Indexed for MEDLINE]

659. Comb Chem High Throughput Screen. 2015;18(4):420-38.

Using Online Tool (iPrior) for Modeling ToxCast™ Assays Towards Prioritization of

Animal Toxicity Testing.

Abdelaziz A, Sushko Y, Novotarskyi S, Körner R, Brandmaier S, Tetko IV(1).

Author information:

(1)eADMET GmbH, Lichtenbergstraße 8, D-85748 Garching/Munich, Germany.

contact@amaziz.com.

The use of long-term animal studies for human and environmental toxicity

estimation is more discouraged than ever before. Alternative models for toxicity

prediction, including QSAR studies, are gaining more ground. A recent approach is

to combine in vitro chemical profiling and in silico chemical descriptors with

the knowledge about toxicity pathways to derive a unique signature for toxicity

endpoints. In this study we investigate the ToxCast™ Phase I data regarding their

ability to predict long-term animal toxicity. We investigated thousands of models

constructed in an effort to predict 61 toxicity endpoints using multiple

descriptor packages and hundreds of in vitro assays. We investigated the use of

in vitro assays and biochemical pathways on model performance. We identified 10

toxicity endpoints where biologically derived descriptors from in vitro assays or

pathway perturbations improved the model prediction ability. In vivo toxicity

endpoints proved generally challenging to model. Few models were possible to

readily model with a balanced accuracy (BA) above 0.7. We also constructed in

silico models to predict the outcome of 144 in vitro assays. This showed better

statistical metrics with 79 out of 144 assays having median balanced accuracy

above 0.7. This suggests that the in vitro datasets have a better modelability

than in vivo animal toxicities for the given datasets. Moreover, we published an

online platform (http://iprior.ochem.eu) that automates large-scale model

building and analysis.

PMID: 25747436 [Indexed for MEDLINE]

660. IEEE Trans Nanobioscience. 2015 Jan;14(1):45-58. doi: 10.1109/TNB.2015.2394328.

Constructing query-driven dynamic machine learning model with application to

protein-ligand binding sites prediction.

Yu DJ, Hu J, Li QM, Tang ZM, Yang JY, Shen HB.

We are facing an era with annotated biological data rapidly and continuously

generated. How to effectively incorporate new annotated data into the learning

step is crucial for enhancing the performance of a bioinformatics prediction

model. Although machine-learning-based methods have been extensively used for

dealing with various biological problems, existing approaches usually train

static prediction models based on fixed training datasets. The static approaches

are found having several disadvantages such as low scalability and impractical

when training dataset is huge. In view of this, we propose a dynamic learning

framework for constructing query-driven prediction models. The key difference

between the proposed framework and the existing approaches is that the training

set for the machine learning algorithm of the proposed framework is dynamically

generated according to the query input, as opposed to training a general model

regardless of queries in traditional static methods. Accordingly, a query-driven

predictor based on the smaller set of data specifically selected from the entire

annotated base dataset will be applied on the query. The new way for constructing

the dynamic model enables us capable of updating the annotated base dataset

flexibly and using the most relevant core subset as the training set makes the

constructed model having better generalization ability on the query, showing

"part could be better than all" phenomenon. According to the new framework, we

have implemented a dynamic protein-ligand binding sites predictor called OSML

(On-site model for ligand binding sites prediction). Computer experiments on 10

different ligand types of three hierarchically organized levels show that OSML

outperforms most existing predictors. The results indicate that the current

dynamic framework is a promising future direction for bridging the gap between

the rapidly accumulated annotated biological data and the effective

machine-learning-based predictors. OSML web server and datasets are freely

available at: http://www.csbio.sjtu.edu.cn/bioinf/OSML/ for academic use.

DOI: 10.1109/TNB.2015.2394328

PMID: 25730499 [Indexed for MEDLINE]

661. J Biomol Struct Dyn. 2015;33(10):2221-33. doi: 10.1080/07391102.2014.998710. Epub

2015 Jan 14.

iDrug-Target: predicting the interactions between drug compounds and target

proteins in cellular networking via benchmark dataset optimization approach.

Xiao X(1), Min JL, Lin WZ, Liu Z, Cheng X, Chou KC.

Author information:

(1)a Computer Department , Jing-De-Zhen Ceramic Institute , Jing-De-Zhen 333046 ,

China.

Information about the interactions of drug compounds with proteins in cellular

networking is very important for drug development. Unfortunately, all the

existing predictors for identifying drug-protein interactions were trained by a

skewed benchmark data-set where the number of non-interactive drug-protein pairs

is overwhelmingly larger than that of the interactive ones. Using this kind of

highly unbalanced benchmark data-set to train predictors would lead to the

outcome that many interactive drug-protein pairs might be mispredicted as

non-interactive. Since the minority interactive pairs often contain the most

important information for drug design, it is necessary to minimize this kind of

misprediction. In this study, we adopted the neighborhood cleaning rule and

synthetic minority over-sampling technique to treat the skewed benchmark datasets

and balance the positive and negative subsets. The new benchmark datasets thus

obtained are called the optimized benchmark datasets, based on which a new

predictor called iDrug-Target was developed that contains four sub-predictors:

iDrug-GPCR, iDrug-Chl, iDrug-Ezy, and iDrug-NR, specialized for identifying the

interactions of drug compounds with GPCRs (G-protein-coupled receptors), ion

channels, enzymes, and NR (nuclear receptors), respectively. Rigorous

cross-validations on a set of experiment-confirmed datasets have indicated that

these new predictors remarkably outperformed the existing ones for the same

purpose. To maximize users' convenience, a public accessible Web server for

iDrug-Target has been established at http://www.jci-bioinfo.cn/iDrug-Target/ , by

which users can easily get their desired results. It has not escaped our notice

that the aforementioned strategy can be widely used in many other areas as well.

DOI: 10.1080/07391102.2014.998710

PMID: 25513722 [Indexed for MEDLINE]

662. J Biomol Struct Dyn. 2015;33(8):1731-42. doi: 10.1080/07391102.2014.968875. Epub

2014 Nov 6.

iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting

sequence evolution information via a gray system model.

Qiu WR(1), Xiao X, Lin WZ, Chou KC.

Author information:

(1)a Computer Department, Jing-De-Zhen Ceramic Institute , Jing-De-Zhen 333403 ,

China.

As one of the most important posttranslational modifications (PTMs),

ubiquitination plays an important role in regulating varieties of biological

processes, such as signal transduction, cell division, apoptosis, and immune

response. Ubiquitination is also named "lysine ubiquitination" because it occurs

when an ubiquitin is covalently attached to lysine (K) residues of targeting

proteins. Given an uncharacterized protein sequence that contains many lysine

residues, which one of them is the ubiquitination site, and which one is of

non-ubiquitination site? With the avalanche of protein sequences generated in the

postgenomic age, it is highly desired for both basic research and drug

development to develop an automated method for rapidly and accurately annotating

the ubiquitination sites in proteins. In view of this, a new predictor called

"iUbiq-Lys" was developed based on the evolutionary information, gray system

model, as well as the general form of pseudo-amino acid composition. It was

demonstrated via the rigorous cross-validations that the new predictor remarkably

outperformed all its counterparts. As a web-server, iUbiq-Lys is accessible to

the public at http://www.jci-bioinfo.cn/iUbiq-Lys . For the convenience of most

experimental scientists, we have further provided a protocol of step-by-step

guide, by which users can easily get their desired results without the need to

follow the complicated mathematics that were presented in this paper just for the

integrity of its development process.

DOI: 10.1080/07391102.2014.968875

PMID: 25248923 [Indexed for MEDLINE]

663. Med Dosw Mikrobiol. 2015;67(2):79-88.

[Differentiation of spa types and staphylococcal cassette chromosome mec (SCCmec)

in clinical methicillin-resistant Staphylococcus aureus isolated in medical sites

of Gdańsk region].

[Article in Polish]

Kasprzyk J, Piechowicz L, Wiśniewska K, Dziewit Ł, Bronk M, Świeć K.

INTRODUCTION: Methicillin-resistant Staphylococcus aureus bacteria are one of the

key etiological factors of hospital-acquired and community-acquired infections.

MRSA strains have an ability of causing a broad spectrum infections: from a

relatively mild skin infections to severe life-threatening systemic infections.

They are characterized by multi-drug resistance, virulence of a number of

factors, may clonally spread within the hospitals and between hospitals.

METHODS: The study embraced a number of 75 isolates of MRSA isolated from

patients of 7 medical sites of the Gdansk region within the period of six months

(June to December 2013). Strains have derived from various clinical materials,

both of hospitalized patients (n=59) and outpatient (n=16). The isolates were

tested for the susceptibility to antimicrobial agents accordance with the

guidelines EUCAST. To estimate of the variability of occurrence of S. aureus

clones used were standard spa gene, consisting in the amplified polymorphic

region of the X gene encoding the protein A gene (spa). After receiving the

results, a spa types were identified using international database Ridom Spa

Server (www.spaserver.ridom.de). To determine the polymorphism cassette carrying

the inecA gene from MRSA strains, used typing five major chromosomal cassette

SCCmec (I-V) by multiplex PCR.

RESULTS: MRSA population genetic analysis carried out on the basis of typing

SCCmec cassettes and spa gene has showed a predominance of strains with SCCmec

type II casette (46.7%) and SCCmec IV casette (38.7%). Less frequently detected

were strains containing SCCmec I cassette (12.0%) and SCCmec III cassette (2.6%).

Spa typing revealed the presence of 13 gene types in MRSA. The most frequently

observed spa types were: t151 (24.0%), t003 (16.0%) in strains of the SCCmec II

cassette and t437 (16.0%) and t008 (14.8%) in the isolates with SCCmec cassette

IV, whereas staphylococcus with the type of spa t011 (12.0%) had SCCmec cassette

I.

CONCLUSIONS: In our population most frequent strains cassette SCCmec II (46.7%),

in most representing types of spa t151 (51.4%) and t003 (34.3%), generally

resistant not only to β-lactam antibiotics, but as erythromycin, clindamycin and

norfloxacin (82.8%), the more frequently they were isolated from patients than a

hospital outpatient centers. The strains SCCmec IV that represent the majority of

outpatient centers (68.8%), the most represented type t437 (41.4%) and often

occurred in hospital centers.

PMID: 26591659 [Indexed for MEDLINE]

664. Methods Mol Biol. 2015;1279:153-65. doi: 10.1007/978-1-4939-2398-4\_10.

A statistical framework for improving genomic annotations of transposon

mutagenesis (TM) assigned essential genes.

Deng J(1).

Author information:

(1)Division of Epidemiology and Biostatistics, Department of Environmental

Health, University of Cincinnati Medical Center, 3223 Eden Av. ML 56, Cincinnati,

OH, 45267-0056, USA, dengjn@mail.uc.edu.

Whole-genome transposon mutagenesis (TM) experiment followed by sequence-based

identification of insertion sites is the most popular genome-wise experiment to

identify essential genes in Prokaryota. However, due to the limitation of

high-throughput technique, this approach yields substantial systematic biases

resulting in the incorrect assignments of many essential genes. To obtain

unbiased and accurate annotations of essential genes from TM experiments, we

developed a novel Poisson model based statistical framework to refine these TM

assignments. In the model, first we identified and incorporated several potential

factors such as gene length and TM insertion information which may cause the TM

assignment biases into the basic Poisson model. Then we calculated the

conditional probability of an essential gene given the observed TM insertion

number. By factorizing this probability through introducing a latent variable the

real insertion number, we formalized the statistical framework. Through

iteratively updating and optimizing model parameters to maximize the

goodness-of-fit of the model to the observed TM insertion data, we finalized the

model. Using this model, we are able to assign the probability score of

essentiality to each individual gene given its TM assignment, which subsequently

correct the experimental biases. To enable our model widely useable, we

established a user-friendly Web-server that is accessible to the public:

http://research.cchmc.org/essentialgene/.

DOI: 10.1007/978-1-4939-2398-4\_10

PMID: 25636618 [Indexed for MEDLINE]

665. Mol Inform. 2015 Jan;34(1):8-17. doi: 10.1002/minf.201400025. Epub 2014 Sep 26.

PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and

Physicochemical Distance Transformation.

Liu B(1,)(2,)(3), Xu J(4), Fan S(4), Xu R(4,)(5), Zhou J(4), Wang X(4,)(5).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong, P.R. China. bliu@insun.hit.edu.cn.

(2)Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute

of Technology Shenzhen Graduate School, Shenzhen, Guangdong, P.R. China.

bliu@insun.hit.edu.cn. (3)Shanghai Key Laboratory of Intelligent Information

Processing, Shanghai, P.R. China. bliu@insun.hit.edu.cn. (4)School of Computer

Science and Technology, Harbin Institute of Technology Shenzhen Graduate School,

Shenzhen, Guangdong, P.R. China. (5)Key Laboratory of Network Oriented

Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School,

Shenzhen, Guangdong, P.R. China.

Identification of DNA-binding proteins is an important problem in biomedical

research as DNA-binding proteins are crucial for various cellular processes.

Currently, the machine learning methods achieve the-state-of-the-art performance

with different features. A key step to improve the performance of these methods

is to find a suitable representation of proteins. In this study, we proposed a

feature vector composed of three kinds of sequence-based features, including

overall amino acid composition, pseudo amino acid composition (PseAAC) proposed

by Chou and physicochemical distance transformation. These features not only

consider the sequence composition of proteins, but also incorporate the

sequence-order information of amino acids in proteins. The feature vectors were

fed into Support Vector Machine (SVM) for DNA-binding protein identification. The

proposed method is called PseDNA-Pro. Experiments on stringent benchmark datasets

and independent test datasets by using the Jackknife test showed that PseDNA-Pro

can achieve an accuracy of higher than 80 %, outperforming several

state-of-the-art methods, including DNAbinder, DNA-Prot, and iDNA-Prot. These

results indicate that the combination of various features for DNA-binding protein

prediction is a suitable approach, and the sequence-order information among

residues in proteins is relative for discrimination. For practical applications,

a web-server of PseDNA-Pro was established, which is available from

http://bioinformatics.hitsz.edu.cn/PseDNA-Pro/.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

DOI: 10.1002/minf.201400025

PMID: 27490858

666. Nucleic Acids Res. 2015 Jan;43(1):74-83. doi: 10.1093/nar/gku1261. Epub 2014 Dec

10.

Comprehensive discovery of DNA motifs in 349 human cells and tissues reveals new

features of motifs.

Zheng Y(1), Li X(2), Hu H(3).

Author information:

(1)Department of Electrical Engineering and Computer Science, University of

Central Florida, Orlando, FL 32816, USA. (2)Burnett School of Biomedical Science,

College of Medicine, University of Central Florida, Orlando, FL 32816, USA

xiaoman@mail.ucf.edu. (3)Department of Electrical Engineering and Computer

Science, University of Central Florida, Orlando, FL 32816, USA

xiaoman@mail.ucf.edu.

Comprehensive motif discovery under experimental conditions is critical for the

global understanding of gene regulation. To generate a nearly complete list of

human DNA motifs under given conditions, we employed a novel approach to de novo

discover significant co-occurring DNA motifs in 349 human DNase I hypersensitive

site datasets. We predicted 845 to 1325 motifs in each dataset, for a total of

2684 non-redundant motifs. These 2684 motifs contained 54.02 to 75.95% of the

known motifs in seven large collections including TRANSFAC. In each dataset, we

also discovered 43 663 to 2 013 288 motif modules, groups of motifs with their

binding sites co-occurring in a significant number of short DNA regions. Compared

with known interacting transcription factors in eight resources, the predicted

motif modules on average included 84.23% of known interacting motifs. We further

showed new features of the predicted motifs, such as motifs enriched in proximal

regions rarely overlapped with motifs enriched in distal regions, motifs enriched

in 5' distal regions were often enriched in 3' distal regions, etc. Finally, we

observed that the 2684 predicted motifs classified the cell or tissue types of

the datasets with an accuracy of 81.29%. The resources generated in this study

are available at http://server.cs.ucf.edu/predrem/.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku1261

PMCID: PMC4288161

PMID: 25505144 [Indexed for MEDLINE]

667. Nucleic Acids Res. 2015 Jan;43(Database issue):D432-8. doi: 10.1093/nar/gku1106.

Epub 2014 Nov 11.

PyIgClassify: a database of antibody CDR structural classifications.

Adolf-Bryfogle J(1), Xu Q(2), North B(2), Lehmann A(2), Dunbrack RL Jr(3).

Author information:

(1)Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue,

Philadelphia, PA 19111, USA Program in Molecular and Cell Biology and Genetics,

Drexel University College of Medicine, 245 N. 15th St. Philadelphia, PA 19102,

USA. (2)Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman

Avenue, Philadelphia, PA 19111, USA. (3)Institute for Cancer Research, Fox Chase

Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA

Roland.Dunbrack@fccc.edu.

Classification of the structures of the complementarity determining regions

(CDRs) of antibodies is critically important for antibody structure prediction

and computational design. We have previously performed a clustering of antibody

CDR conformations and defined a systematic nomenclature consisting of the CDR,

length and an integer starting from the largest to the smallest cluster in the

data set (e.g. L1-11-1). We present PyIgClassify (for Python-based immunoglobulin

classification; available at http://dunbrack2.fccc.edu/pyigclassify/), a database

and web server that provides access to assignments of all CDR structures in the

PDB to our classification system. The database includes assignments to the IMGT

germline V regions for heavy and light chains for several species. For humanized

antibodies, the assignment of the frameworks is to human germlines and the CDRs

to the germlines of mice or other species sources. The database can be searched

by PDB entry, cluster identifier and IMGT germline group (e.g. human IGHV1). The

entire database is downloadable so that users may filter the data as needed for

antibody structure analysis, prediction and design.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku1106

PMCID: PMC4383924

PMID: 25392411 [Indexed for MEDLINE]

668. Nucleic Acids Res. 2015 Jan;43(Database issue):D465-9. doi: 10.1093/nar/gku1088.

Epub 2014 Nov 6.

Recent improvements to Binding MOAD: a resource for protein-ligand binding

affinities and structures.

Ahmed A(1), Smith RD(1), Clark JJ(1), Dunbar JB Jr(1), Carlson HA(2).

Author information:

(1)Department of Medicinal Chemistry, University of Michigan, 428 Church St, Ann

Arbor, MI 48109-1065, USA. (2)Department of Medicinal Chemistry, University of

Michigan, 428 Church St, Ann Arbor, MI 48109-1065, USA carlsonh@umich.edu.

For over 10 years, Binding MOAD (Mother of All Databases;

http://www.BindingMOAD.org) has been one of the largest resources for

high-quality protein-ligand complexes and associated binding affinity data.

Binding MOAD has grown at the rate of 1994 complexes per year, on average.

Currently, it contains 23,269 complexes and 8156 binding affinities. Our annual

updates curate the data using a semi-automated literature search of the

references cited within the PDB file, and we have recently upgraded our website

and added new features and functionalities to better serve Binding MOAD users. In

order to eliminate the legacy application server of the old platform and to

accommodate new changes, the website has been completely rewritten in the LAMP

(Linux, Apache, MySQL and PHP) environment. The improved user interface

incorporates current third-party plugins for better visualization of protein and

ligand molecules, and it provides features like sorting, filtering and filtered

downloads. In addition to the field-based searching, Binding MOAD now can be

searched by structural queries based on the ligand. In order to remove

redundancy, Binding MOAD records are clustered in different families based on 90%

sequence identity. The new Binding MOAD, with the upgraded platform, features and

functionalities, is now equipped to better serve its users.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku1088

PMCID: PMC4383918

PMID: 25378330 [Indexed for MEDLINE]

669. Nucleic Acids Res. 2015 Jan;43(Database issue):D682-9. doi: 10.1093/nar/gku1112.

Epub 2014 Nov 6.

Genomicus update 2015: KaryoView and MatrixView provide a genome-wide perspective

to multispecies comparative genomics.

Louis A(1), Nguyen NT(2), Muffato M(3), Roest Crollius H(2).

Author information:

(1)Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris F-75005,

France Inserm U1024, Paris F-75005, France CNRS, UMR 8197, Paris F-75005, France

alouis@biologie.ens.fr. (2)Ecole Normale Supérieure, Institut de Biologie de

l'ENS, IBENS, Paris F-75005, France Inserm U1024, Paris F-75005, France CNRS, UMR

8197, Paris F-75005, France. (3)European Molecular Biology Laboratory, European

Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10

1SD, UK.

The Genomicus web server (http://www.genomicus.biologie.ens.fr/genomicus) is a

visualization tool allowing comparative genomics in four different phyla

(Vertebrate, Fungi, Metazoan and Plants). It provides access to genomic

information from extant species, as well as ancestral gene content and gene order

for vertebrates and flowering plants. Here we present the new features available

for vertebrate genome with a focus on new graphical tools. The interface to enter

the database has been improved, two pairwise genome comparison tools are now

available (KaryoView and MatrixView) and the multiple genome comparison tools

(PhyloView and AlignView) propose three new kinds of representation and a more

intuitive menu. These new developments have been implemented for Genomicus portal

dedicated to vertebrates. This allows the analysis of 68 extant animal genomes,

as well as 58 ancestral reconstructed genomes. The Genomicus server also provides

access to ancestral gene orders, to facilitate evolutionary and comparative

genomics studies, as well as computationally predicted regulatory interactions,

thanks to the representation of conserved non-coding elements with their putative

gene targets.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku1112

PMCID: PMC4383929

PMID: 25378326 [Indexed for MEDLINE]

670. Nucleic Acids Res. 2015 Jan;43(Database issue):D146-52. doi: 10.1093/nar/gku1104.

Epub 2014 Nov 5.

miRDB: an online resource for microRNA target prediction and functional

annotations.

Wong N(1), Wang X(2).

Author information:

(1)Department of Biomedical Engineering, Washington University, St. Louis, MO

63130, USA Department of Radiation Oncology, Washington University School of

Medicine, St. Louis, MO 63108, USA. (2)Department of Radiation Oncology,

Washington University School of Medicine, St. Louis, MO 63108, USA Department of

Biomedical Engineering, Washington University, St. Louis, MO 63130, USA

xwang@radonc.wustl.edu.

MicroRNAs (miRNAs) are small non-coding RNAs that are extensively involved in

many physiological and disease processes. One major challenge in miRNA studies is

the identification of genes regulated by miRNAs. To this end, we have developed

an online resource, miRDB (http://mirdb.org), for miRNA target prediction and

functional annotations. Here, we describe recently updated features of miRDB,

including 2.1 million predicted gene targets regulated by 6709 miRNAs. In

addition to presenting precompiled prediction data, a new feature is the web

server interface that allows submission of user-provided sequences for miRNA

target prediction. In this way, users have the flexibility to study any custom

miRNAs or target genes of interest. Another major update of miRDB is related to

functional miRNA annotations. Although thousands of miRNAs have been identified,

many of the reported miRNAs are not likely to play active functional roles or may

even have been falsely identified as miRNAs from high-throughput studies. To

address this issue, we have performed combined computational analyses and

literature mining, and identified 568 and 452 functional miRNAs in humans and

mice, respectively. These miRNAs, as well as associated functional annotations,

are presented in the FuncMir Collection in miRDB.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku1104

PMCID: PMC4383922

PMID: 25378301 [Indexed for MEDLINE]

671. Nucleic Acids Res. 2015 Jan;43(Database issue):D996-1002. doi:

10.1093/nar/gku1053. Epub 2014 Oct 29.

AraNet v2: an improved database of co-functional gene networks for the study of

Arabidopsis thaliana and 27 other nonmodel plant species.

Lee T(1), Yang S(1), Kim E(1), Ko Y(1), Hwang S(2), Shin J(1), Shim JE(1), Shim

H(1), Kim H(1), Kim C(1), Lee I(3).

Author information:

(1)Department of Biotechnology, College of Life Science and Biotechnology, Yonsei

University, Seoul, Korea. (2)Department of Biotechnology, College of Life Science

and Biotechnology, Yonsei University, Seoul, Korea Center for Systems and

Synthetic Biology, Institute for Cellular and Molecular Biology, University of

Texas at Austin, TX 78712, USA. (3)Department of Biotechnology, College of Life

Science and Biotechnology, Yonsei University, Seoul, Korea insuklee@yonsei.ac.kr.

Arabidopsis thaliana is a reference plant that has been studied intensively for

several decades. Recent advances in high-throughput experimental technology have

enabled the generation of an unprecedented amount of data from A. thaliana, which

has facilitated data-driven approaches to unravel the genetic organization of

plant phenotypes. We previously published a description of a genome-scale

functional gene network for A. thaliana, AraNet, which was constructed by

integrating multiple co-functional gene networks inferred from diverse data

types, and we demonstrated the predictive power of this network for complex

phenotypes. More recently, we have observed significant growth in the

availability of omics data for A. thaliana as well as improvements in data

analysis methods that we anticipate will further enhance the integrated database

of co-functional networks. Here, we present an updated co-functional gene network

for A. thaliana, AraNet v2 (available at http://www.inetbio.org/aranet), which

covers approximately 84% of the coding genome. We demonstrate significant

improvements in both genome coverage and accuracy. To enhance the usability of

the network, we implemented an AraNet v2 web server, which generates functional

predictions for A. thaliana and 27 nonmodel plant species using an

orthology-based projection of nonmodel plant genes on the A. thaliana gene

network.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku1053

PMCID: PMC4383895

PMID: 25355510 [Indexed for MEDLINE]

672. Nucleic Acids Res. 2015 Jan;43(Database issue):D662-9. doi: 10.1093/nar/gku1010.

Epub 2014 Oct 28.

Ensembl 2015.

Cunningham F(1), Amode MR(1), Barrell D(2), Beal K(1), Billis K(1), Brent S(3),

Carvalho-Silva D(1), Clapham P(3), Coates G(3), Fitzgerald S(1), Gil L(1), Girón

CG(1), Gordon L(1), Hourlier T(1), Hunt SE(1), Janacek SH(1), Johnson N(1),

Juettemann T(1), Kähäri AK(3), Keenan S(1), Martin FJ(1), Maurel T(1), McLaren

W(1), Murphy DN(2), Nag R(1), Overduin B(1), Parker A(1), Patricio M(1), Perry

E(1), Pignatelli M(1), Riat HS(1), Sheppard D(1), Taylor K(1), Thormann A(1),

Vullo A(1), Wilder SP(1), Zadissa A(1), Aken BL(1), Birney E(1), Harrow J(3),

Kinsella R(1), Muffato M(1), Ruffier M(1), Searle SM(3), Spudich G(1), Trevanion

SJ(1), Yates A(1), Zerbino DR(1), Flicek P(4).

Author information:

(1)European Molecular Biology Laboratory, European Bioinformatics Institute,

Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. (2)European

Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust

Genome Campus, Hinxton, Cambridge CB10 1SD, UK Wellcome Trust Sanger Institute,

Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. (3)Wellcome Trust

Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.

(4)European Molecular Biology Laboratory, European Bioinformatics Institute,

Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK Wellcome Trust

Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

flicek@ebi.ac.uk.

Ensembl (http://www.ensembl.org) is a genomic interpretation system providing the

most up-to-date annotations, querying tools and access methods for chordates and

key model organisms. This year we released updated annotation (gene models,

comparative genomics, regulatory regions and variation) on the new human

assembly, GRCh38, although we continue to support researchers using the

GRCh37.p13 assembly through a dedicated site (http://grch37.ensembl.org). Our

Regulatory Build has been revamped to identify regulatory regions of interest and

to efficiently highlight their activity across disparate epigenetic data sets. A

number of new interfaces allow users to perform large-scale comparisons of their

data against our annotations. The REST server (http://rest.ensembl.org), which

allows programs written in any language to query our databases, has moved to a

full service alongside our upgraded website tools. Our online Variant Effect

Predictor tool has been updated to process more variants and calculate summary

statistics. Lastly, the WiggleTools package enables users to summarize large

collections of data sets and view them as single tracks in Ensembl. The Ensembl

code base itself is more accessible: it is now hosted on our GitHub organization

page (https://github.com/Ensembl) under an Apache 2.0 open source license.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku1010

PMCID: PMC4383879

PMID: 25352552 [Indexed for MEDLINE]

673. Nucleic Acids Res. 2015 Jan;43(Database issue):D76-81. doi: 10.1093/nar/gku887.

Epub 2014 Sep 27.

AnimalTFDB 2.0: a resource for expression, prediction and functional study of

animal transcription factors.

Zhang HM(1), Liu T(1), Liu CJ(1), Song S(1), Zhang X(1), Liu W(1), Jia H(1), Xue

Y(1), Guo AY(2).

Author information:

(1)Department of Biomedical Engineering, Key Laboratory of Molecular Biophysics

of the Ministry of Education, College of Life Science and Technology, Huazhong

University of Science and Technology, Wuhan, Hubei 430074, PR China.

(2)Department of Biomedical Engineering, Key Laboratory of Molecular Biophysics

of the Ministry of Education, College of Life Science and Technology, Huazhong

University of Science and Technology, Wuhan, Hubei 430074, PR China

guoay@mail.hust.edu.cn.

Transcription factors (TFs) are key regulators for gene expression. Here we

updated the animal TF database AnimalTFDB to version 2.0

(http://bioinfo.life.hust.edu.cn/AnimalTFDB/). Using the improved prediction

pipeline, we identified 72 336 TF genes, 21 053 transcription co-factor genes and

6502 chromatin remodeling factor genes from 65 species covering main animal

lineages. Besides the abundant annotations (basic information, gene model,

protein functional domain, gene ontology, pathway, protein interaction, ortholog

and paralog, etc.) in the previous version, we made several new features and

functions in the updated version. These new features are: (i) gene expression

from RNA-Seq for nine model species, (ii) gene phenotype information, (iii)

multiple sequence alignment of TF DNA-binding domains, and the weblogo and

phylogenetic tree based on the alignment, (iv) a TF prediction server to identify

new TFs from input sequences and (v) a BLAST server to search against TFs in

AnimalTFDB. A new nice web interface was designed for AnimalTFDB 2.0 allowing

users to browse and search all data in the database. We aim to maintain the

AnimalTFDB as a solid resource for TF identification and studies of transcription

regulation and comparative genomics.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku887

PMCID: PMC4384004

PMID: 25262351 [Indexed for MEDLINE]

674. Protein Sci. 2015 Jan;24(1):145-53. doi: 10.1002/pro.2581. Epub 2014 Oct 25.

Use of a structural alphabet to find compatible folds for amino acid sequences.

Mahajan S(1), de Brevern AG, Sanejouand YH, Srinivasan N, Offmann B.

Author information:

(1)Université de La Réunion, DSIMB, UMR-S S1134, Saint Denis Messag Cedex 09, La

Réunion, F-97715, France; INSERM, UMR-S 1134, DSIMB, F-75739, Paris, France;

Laboratoire d'Excellence, GR-Ex, Paris, F-75739, France; Université de Nantes,

UFIP CNRS UMR 6286 Faculté des Sciences et Techniques, 2 rue de la Houssinière,

44392, Nantes Cedex 03, France.

The structural annotation of proteins with no detectable homologs of known 3D

structure identified using sequence-search methods is a major challenge today. We

propose an original method that computes the conditional probabilities for the

amino-acid sequence of a protein to fit to known protein 3D structures using a

structural alphabet, known as "Protein Blocks" (PBs). PBs constitute a library of

16 local structural prototypes that approximate every part of protein backbone

structures. It is used to encode 3D protein structures into 1D PB sequences and

to capture sequence to structure relationships. Our method relies on amino acid

occurrence matrices, one for each PB, to score global and local threading of

query amino acid sequences to protein folds encoded into PB sequences. It does

not use any information from residue contacts or sequence-search methods or

explicit incorporation of hydrophobic effect. The performance of the method was

assessed with independent test datasets derived from SCOP 1.75A. With a Z-score

cutoff that achieved 95% specificity (i.e., less than 5% false positives), global

and local threading showed sensitivity of 64.1% and 34.2%, respectively. We

further tested its performance on 57 difficult CASP10 targets that had no known

homologs in PDB: 38 compatible templates were identified by our approach and 66%

of these hits yielded correctly predicted structures. This method scales-up well

and offers promising perspectives for structural annotations at genomic level. It

has been implemented in the form of a web-server that is freely available at

http://www.bo-protscience.fr/forsa.

© 2014 The Protein Society.

DOI: 10.1002/pro.2581

PMCID: PMC4282420

PMID: 25297700 [Indexed for MEDLINE]

675. Stroke. 2015 Jan;46(1):137-42. doi: 10.1161/STROKEAHA.114.007124. Epub 2014 Nov

13.

Feasibility platform for stroke studies: an online tool to improve eligibility

criteria for clinical trials.

Minnerup J(1), Trinczek B(2), Storck M(2), Hohenberger M(2), Wilpsbäumer S(2),

Abdul-Rahim AH(2), Liesirova KT(2), Broeg-Morvay A(2), Meisterernst J(2), Lees

KR(2), Mattle HP(2), Wersching H(2), Dziewas R(2), Schäbitz WR(2), Schilling

M(2).

Author information:

(1)From the Department of Neurology (J. Minnerup, M.H., R.D., M. Schilling),

Institute of Medical Informatics (B.T., M. Storck, S.W.), Institute of

Epidemiology and Social Medicine (H.W.), University of Münster,

Albert-Schweitzer-Campus 1, Münster, Germany; Institute of Cardiovascular and

Medical Sciences, University of Glasgow, Glasgow, United Kingdom (A.H.A.-R.,

K.R.L.); Department of Neurology, Inselspital, University of Bern, Bern,

Switzerland (K.T.L., A.B.-M., J. Meisterernst, H.P.M.); and Department of

Neurology, Bethel, Bielefeld, Germany (W.-R.S.). minnerup@uni-muenster.de.

(2)From the Department of Neurology (J. Minnerup, M.H., R.D., M. Schilling),

Institute of Medical Informatics (B.T., M. Storck, S.W.), Institute of

Epidemiology and Social Medicine (H.W.), University of Münster,

Albert-Schweitzer-Campus 1, Münster, Germany; Institute of Cardiovascular and

Medical Sciences, University of Glasgow, Glasgow, United Kingdom (A.H.A.-R.,

K.R.L.); Department of Neurology, Inselspital, University of Bern, Bern,

Switzerland (K.T.L., A.B.-M., J. Meisterernst, H.P.M.); and Department of

Neurology, Bethel, Bielefeld, Germany (W.-R.S.).

BACKGROUND AND PURPOSE: Eligibility criteria are a key factor for the feasibility

and validity of clinical trials. We aimed to develop an online tool to assess the

potential effect of inclusion and exclusion criteria on the proportion of

patients eligible for an acute stroke trial.

METHODS: We identified relevant inclusion and exclusion criteria of acute stroke

trials. Based on these criteria and using a cohort of 1537 consecutive patients

with acute ischemic stroke from 3 stroke centers, we developed a web portal

feasibility platform for stroke studies (FePASS) to estimate proportions of

eligible patients for acute stroke trials. We applied the FePASS resource to

calculate the proportion of patients eligible for 4 recent stroke studies.

RESULTS: Sixty-one eligibility criteria were derived from 30 trials on acute

ischemic stroke. FePASS, publicly available at http://fepass.uni-muenster.de,

displays the proportion of patients in percent to assess the effect of varying

values of relevant eligibility criteria, for example, age, symptom onset time,

National Institutes of Health Stroke Scale, and prestroke modified Rankin Scale,

on this proportion. The proportion of eligible patients for 4 recent stroke

studies ranged from 2.1% to 11.3%. Slight variations of the inclusion criteria

could substantially increase the proportion of eligible patients.

CONCLUSIONS: FePASS is an open access online resource to assess the effect of

inclusion and exclusion criteria on the proportion of eligible patients for a

stroke trial. FePASS can help to design stroke studies, optimize eligibility

criteria, and to estimate the potential recruitment rate.

© 2014 American Heart Association, Inc.

DOI: 10.1161/STROKEAHA.114.007124

PMID: 25395412 [Indexed for MEDLINE]

676. Stud Health Technol Inform. 2015;212:23-6.

Standardized mappings--a framework to combine different semantic mappers into a

standardized web-API.

Neuhaus P(1), Doods J(1), Dugas M(1).

Author information:

(1)Institute of Medical Informatics, Westfälische Wilhelms-Universität, Münster,

Germany.

BACKGROUND: Automatic coding of medical terms is an important, but highly

complicated and laborious task.

OBJECTIVES: To compare and evaluate different strategies a framework with a

standardized web-interface was created. Two UMLS mapping strategies are compared

to demonstrate the interface.

METHODS: The framework is a Java Spring application running on a Tomcat

application server. It accepts different parameters and returns results in JSON

format. To demonstrate the framework, a list of medical data items was mapped by

two different methods: similarity search in a large table of terminology codes

versus search in a manually curated repository. These mappings were reviewed by a

specialist.

RESULTS: The evaluation shows that the framework is flexible (due to standardized

interfaces like HTTP and JSON), performant and reliable. Accuracy of

automatically assigned codes is limited (up to 40%).

CONCLUSION: Combining different semantic mappers into a standardized Web-API is

feasible. This framework can be easily enhanced due to its modular design.

PMID: 26063253 [Indexed for MEDLINE]

677. BMC Bioinformatics. 2014 Dec 30;15:427. doi: 10.1186/s12859-014-0427-6.

WEBnm@ v2.0: Web server and services for comparing protein flexibility.

Tiwari SP(1,)(2), Fuglebakk E(3), Hollup SM(4), Skjærven L(5,)(6), Cragnolini

T(7,)(8,)(9), Grindhaug SH(10), Tekle KM(11), Reuter N(12,)(13).

Author information:

(1)Department of Molecular Biology, University of Bergen, Bergen, Norway.

Sandhya.Tiwari@mbi.uib.no. (2)Computational Biology Unit, Department of

Informatics, University of Bergen, Bergen, Norway. Sandhya.Tiwari@mbi.uib.no.

(3)Computational Biology Unit, Department of Informatics, University of Bergen,

Bergen, Norway. Edvin.Fuglebakk@mbi.uib.no. (4)Computational Biology Unit,

Department of Informatics, University of Bergen, Bergen, Norway.

siv.hollup@gmail.com. (5)Department of Biomedicine, University of Bergen, Bergen,

Norway. Lars.Skjarven@biomed.uib.no. (6)Computational Biology Unit, Department of

Informatics, University of Bergen, Bergen, Norway. Lars.Skjarven@biomed.uib.no.

(7)Department of Molecular Biology, University of Bergen, Bergen, Norway.

tristan.cragnolini@gmail.com. (8)Computational Biology Unit, Department of

Informatics, University of Bergen, Bergen, Norway. tristan.cragnolini@gmail.com.

(9)Present address: University Chemical Laboratories, University of Cambridge,

Lensfield Road, Cambridge, CB2 1EW, UK. tristan.cragnolini@gmail.com.

(10)Computational Biology Unit, Department of Informatics, University of Bergen,

Bergen, Norway. svennhg@gmail.com. (11)Computational Biology Unit, Department of

Informatics, University of Bergen, Bergen, Norway. kidane.tekle@cbu.uib.no.

(12)Department of Molecular Biology, University of Bergen, Bergen, Norway.

Nathalie.Reuter@mbi.uib.no. (13)Computational Biology Unit, Department of

Informatics, University of Bergen, Bergen, Norway. Nathalie.Reuter@mbi.uib.no.

BACKGROUND: Normal mode analysis (NMA) using elastic network models is a reliable

and cost-effective computational method to characterise protein flexibility and

by extension, their dynamics. Further insight into the dynamics-function

relationship can be gained by comparing protein motions between protein homologs

and functional classifications. This can be achieved by comparing normal modes

obtained from sets of evolutionary related proteins.

RESULTS: We have developed an automated tool for comparative NMA of a set of

pre-aligned protein structures. The user can submit a sequence alignment in the

FASTA format and the corresponding coordinate files in the Protein Data Bank

(PDB) format. The computed normalised squared atomic fluctuations and atomic

deformation energies of the submitted structures can be easily compared on graphs

provided by the web user interface. The web server provides pairwise comparison

of the dynamics of all proteins included in the submitted set using two measures:

the Root Mean Squared Inner Product and the Bhattacharyya Coefficient. The

Comparative Analysis has been implemented on our web server for NMA, WEBnm@,

which also provides recently upgraded functionality for NMA of single protein

structures. This includes new visualisations of protein motion, visualisation of

inter-residue correlations and the analysis of conformational change using the

overlap analysis. In addition, programmatic access to WEBnm@ is now available

through a SOAP-based web service. Webnm@ is available at

http://apps.cbu.uib.no/webnma .

CONCLUSION: WEBnm@ v2.0 is an online tool offering unique capability for

comparative NMA on multiple protein structures. Along with a convenient web

interface, powerful computing resources, and several methods for mode analyses,

WEBnm@ facilitates the assessment of protein flexibility within protein families

and superfamilies. These analyses can give a good view of how the structures move

and how the flexibility is conserved over the different structures.

DOI: 10.1186/s12859-014-0427-6

PMCID: PMC4339738

PMID: 25547242 [Indexed for MEDLINE]

678. Bioinform Biol Insights. 2014 Dec 21;8:209-14. doi: 10.4137/BBI.S19057.

eCollection 2014.

POEAS: Automated Plant Phenomic Analysis Using Plant Ontology.

Shameer K(1), Naika MB(2), Mathew OK(1), Sowdhamini R(1).

Author information:

(1)National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore, India.

(2)National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore, India.

; Department of Plant Biotechnology, University of Agricultural Sciences, GKVK

Campus, Bangalore, India.

Biological enrichment analysis using gene ontology (GO) provides a global

overview of the functional role of genes or proteins identified from large-scale

genomic or proteomic experiments. Phenomic enrichment analysis of gene lists can

provide an important layer of information as well as cellular components,

molecular functions, and biological processes associated with gene lists. Plant

phenomic enrichment analysis will be useful for performing new experiments to

better understand plant systems and for the interpretation of gene or proteins

identified from high-throughput experiments. Plant ontology (PO) is a compendium

of terms to define the diverse phenotypic characteristics of plant species,

including plant anatomy, morphology, and development stages. Adoption of this

highly useful ontology is limited, when compared to GO, because of the lack of

user-friendly tools that enable the use of PO for statistical enrichment

analysis. To address this challenge, we introduce Plant Ontology Enrichment

Analysis Server (POEAS) in the public domain. POEAS uses a simple list of genes

as input data and performs enrichment analysis using Ontologizer 2.0 to provide

results in two levels, enrichment results and visualization utilities, to

generate ontological graphs that are of publication quality. POEAS also offers

interactive options to identify user-defined background population sets, various

multiple-testing correction methods, different enrichment calculation methods,

and resampling tests to improve statistical significance. The availability of

such a tool to perform phenomic enrichment analyses using plant genes as a

complementary resource will permit the adoption of PO-based phenomic analysis as

part of analytical workflows. POEAS can be accessed using the URL

http://caps.ncbs.res.in/poeas.

DOI: 10.4137/BBI.S19057

PMCID: PMC4274039

PMID: 25574136

679. J Theor Biol. 2014 Dec 21;363:412-8. doi: 10.1016/j.jtbi.2014.08.002. Epub 2014

Aug 11.

PECM: prediction of extracellular matrix proteins using the concept of Chou's

pseudo amino acid composition.

Zhang J(1), Sun P(2), Zhao X(1), Ma Z(3).

Author information:

(1)School of Computer Science and Information Technology, Northeast Normal

University, Changchun 1300117, PR China. (2)School of Computer Science and

Information Technology, Northeast Normal University, Changchun 1300117, PR China;

National Engineering Laboratory for Druggable Gene and Protein Screening,

Northeast Normal University, Changchun 130024, PR China. (3)School of Computer

Science and Information Technology, Northeast Normal University, Changchun

1300117, PR China. Electronic address: zhiqiangma.nenu@gmail.com.

The extracellular matrix proteins (ECMs) are widely found in the tissues of

multicellular organisms. They consist of various secreted proteins, mainly

polysaccharides and glycoproteins. The ECMs involve the exchange of materials and

information between resident cells and the external environment. Accurate

identification of ECMs is a significant step in understanding the evolution of

cancer as well as promises wide range of potential applications in therapeutic

targets or diagnostic markers. In this paper, an accurate computational method

named PECM is proposed for identifying ECMs. Here, we explore various

sequence-derived discriminative features including evolutionary information,

predicted secondary structure, and physicochemical properties. Rather than simply

combining the features which may bring information redundancy and unwanted

noises, we use Fisher-Markov selector and incremental feature selection approach

to search the optimal feature subsets. Then, we train our model by the technique

of support vector machine (SVM). PECM achieves good prediction performance with

the ACC scores about 86% and 90% on testing and independent datasets, which are

competitive with the state-of-the-art ECMs prediction tools. A web-server named

PECM which implements the proposed approach is freely available at

http://59.73.198.144:8088/PECM/.

Copyright © 2014 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2014.08.002

PMID: 25123433 [Indexed for MEDLINE]

680. BMC Bioinformatics. 2014 Dec 19;15:414. doi: 10.1186/s12859-014-0414-y.

EPMLR: sequence-based linear B-cell epitope prediction method using multiple

linear regression.

Lian Y(1), Ge M(2), Pan XM(3).

Author information:

(1)The Key Laboratory of Bioinformatics, Ministry of Education, School of Life

Sciences, Tsinghua University, Beijing, 100084, China. lianyao1112@gmail.com.

(2)CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of

Genomics, Chinese Academy of Sciences, Beijing, China. gemeng@big.ac.cn. (3)The

Key Laboratory of Bioinformatics, Ministry of Education, School of Life Sciences,

Tsinghua University, Beijing, 100084, China. pan-xm@mail.tsinghua.edu.cn.

BACKGROUND: B-cell epitopes have been studied extensively due to their

immunological applications, such as peptide-based vaccine development, antibody

production, and disease diagnosis and therapy. Despite several decades of

research, the accurate prediction of linear B-cell epitopes has remained a

challenging task.

RESULTS: In this work, based on the antigen's primary sequence information, a

novel linear B-cell epitope prediction model was developed using the multiple

linear regression (MLR). A 10-fold cross-validation test on a large non-redundant

dataset was performed to evaluate the performance of our model. To alleviate the

problem caused by the noise of negative dataset, 300 experiments utilizing 300

sub-datasets were performed. We achieved overall sensitivity of 81.8%, precision

of 64.1% and area under the receiver operating characteristic curve (AUC) of

0.728.

CONCLUSIONS: We have presented a reliable method for the identification of linear

B cell epitope using antigen's primary sequence information. Moreover, a web

server EPMLR has been developed for linear B-cell epitope prediction:

http://www.bioinfo.tsinghua.edu.cn/epitope/EPMLR/ .

DOI: 10.1186/s12859-014-0414-y

PMCID: PMC4307399

PMID: 25523327 [Indexed for MEDLINE]

681. BMC Plant Biol. 2014 Dec 18;14:315. doi: 10.1186/s12870-014-0315-2.

The chickpea genomic web resource: visualization and analysis of the desi-type

Cicer arietinum nuclear genome for comparative exploration of legumes.

Misra G, Priya P, Bandhiwal N, Bareja N, Jain M, Bhatia S, Chattopadhyay D, Tyagi

AK, Yadav G.

BACKGROUND: Availability of the draft nuclear genome sequences of small-seeded

desi-type legume crop Cicer arietinum has provided an opportunity for

investigating unique chickpea genomic features and evaluation of their biological

significance. The increasing number of legume genome sequences also presents a

challenge for developing reliable and information-driven bioinformatics

applications suitable for comparative exploration of this important class of crop

plants.

RESULTS: The Chickpea Genomic Web Resource (CGWR) is an implementation of a suite

of web-based applications dedicated to chickpea genome visualization and

comparative analysis, based on next generation sequencing and assembly of Cicer

arietinum desi-type genotype ICC4958. CGWR has been designed and configured for

mapping, scanning and browsing the significant chickpea genomic features in view

of the important existing and potential roles played by the various legume genome

projects in mutant mapping and cloning. It also enables comparative informatics

of ICC4958 DNA sequence analysis with other wild and cultivated genotypes of

chickpea, various other leguminous species as well as several non-leguminous

model plants, to enable investigations into evolutionary processes that shape

legume genomes.

CONCLUSIONS: CGWR is an online database offering a comprehensive visual and

functional genomic analysis of the chickpea genome, along with customized maps

and gene-clustering options. It is also the only plant based web resource

supporting display and analysis of nucleosome positioning patterns in the genome.

The usefulness of CGWR has been demonstrated with discoveries of biological

significance made using this server. The CGWR is compatible with all available

operating systems and browsers, and is available freely under the open source

license at http://www.nipgr.res.in/CGWR/home.php.

DOI: 10.1186/s12870-014-0315-2

PMCID: PMC4307184

PMID: 25518738 [Indexed for MEDLINE]

682. Algorithms Mol Biol. 2014 Dec 16;9(1):25. doi: 10.1186/s13015-014-0025-1.

eCollection 2014.

Analysis of pattern overlaps and exact computation of P-values of pattern

occurrences numbers: case of Hidden Markov Models.

Régnier M(1), Furletova E(2), Yakovlev V(3), Roytberg M(4).

Author information:

(1)INRIA, d'Estienne d'Orves 1, Palaiseau, 91120 France ; CNRS, d'Estienne

d'Orves 1, Palaiseau, 91120 France ; LIX, Ecole Polytechnique, Batiment A.

Turing, Palaiseau, 91128 France. (2)Institute of Mathematical Problems of

Biology, 142290, Institutskaya, 4, Pushchino, Russia ; Pushchino State

University, 142290, Prospect Nauki, 5, Pushchino, Russia. (3)Institute of

Mathematical Problems of Biology, 142290, Institutskaya, 4, Pushchino, Russia ;

National Research University "Higher School of Economics", 101978, Myasnitskaya

str., 20, Moscow, Russia. (4)Institute of Mathematical Problems of Biology,

142290, Institutskaya, 4, Pushchino, Russia ; Laboratoire J.-V. Poncelet (UMI

2615), 119002, Bolshoy Vlasyevskiy Pereulok, 11, Moscow, Russia ; National

Research University "Higher School of Economics", 101978, Myasnitskaya str., 20,

Moscow, Russia ; Moscow Institute of Physics and Technology, 141700, Institutsky

pereulok, 11, Moscow Region, Dolgoprudny, Russia.

BACKGROUND: Finding new functional fragments in biological sequences is a

challenging problem. Methods addressing this problem commonly search for clusters

of pattern occurrences that are statistically significant. A measure of

statistical significance is the P-value of a number of pattern occurrences, i.e.

the probability to find at least S occurrences of words from a pattern in a

random text of length N generated according to a given probability model. All

words of the pattern are supposed to be of same length.

RESULTS: We present a novel algorithm SufPref that computes an exact P-value for

Hidden Markov models (HMM). The algorithm is based on recursive equations on text

sets related to pattern occurrences; the equations can be used for any

probability model. The algorithm inductively traverses a specific data structure,

an overlap graph. The nodes of the graph are associated with the overlaps of

words from . The edges are associated to the prefix and suffix relations between

overlaps. An originality of our data structure is that pattern need not be

explicitly represented in nodes or leaves. The algorithm relies on the Cartesian

product of the overlap graph and the graph of HMM states; this approach is

analogous to the automaton approach from JBCB 4: 553-569. The gain in size of

SufPref data structure leads to significant improvements in space and time

complexity compared to existent algorithms. The algorithm SufPref was implemented

as a C++ program; the program can be used both as Web-server and a stand alone

program for Linux and Windows. The program interface admits special formats to

describe probability models of various types (HMM, Bernoulli, Markov); a pattern

can be described with a list of words, a PSSM, a degenerate pattern or a word and

a number of mismatches. It is available at

http://server2.lpm.org.ru/bio/online/sf/. The program was applied to compare

sensitivity and specificity of methods for TFBS prediction based on P-values

computed for Bernoulli models, Markov models of orders one and two and HMMs. The

experiments show that the methods have approximately the same qualities.

DOI: 10.1186/s13015-014-0025-1

PMCID: PMC4307674

PMID: 25648087

683. Sci Rep. 2014 Dec 15;4:7476. doi: 10.1038/srep07476.

A hybrid method for identification of structural domains.

Hua Y(1), Zhu M(1), Wang Y(2), Xie Z(1), Li M(2).

Author information:

(1)College of Computer Science, Sichuan University, No.24 South Section 1, Yihuan

Road, 610064 Chengdu, China. (2)College of Chemistry, Sichuan University, No.24

South Section 1, Yihuan Road, 610065 Chengdu, China.

Structural domains in proteins are the basic units to form various proteins. In

the protein's evolution and functioning, domains play important roles. But the

definition of domain is not yet precisely given, and the update cycle of

structural domain databases is long. The automatic algorithms identify domains

slowly, while protein entities with great structural complexity are on the rise.

Here, we present a method which recognizes the compact and modular segments of

polypeptide chains to identify structural domains, and contrast some data sets to

illuminate their effect. The method combines support vector machine (SVM) with

K-means algorithm. It is faster and more stable than most current algorithms and

performs better. It also indicates that when proteins are presented as some

Alpha-carbon atoms in 3D space, it is feasible to identify structural domains by

the spatially structural properties. We have developed a web-server, which would

be helpful in identification of structural domains

(http://vis.sculab.org/~huayongpan/cgi-bin/domainAssignment.cgi).

DOI: 10.1038/srep07476

PMCID: PMC4265785

PMID: 25503992 [Indexed for MEDLINE]

684. BMC Bioinformatics. 2014 Dec 11;15:408. doi: 10.1186/s12859-014-0408-9.

Kiwi: a tool for integration and visualization of network topology and gene-set

analysis.

Väremo L, Gatto F, Nielsen J(1).

Author information:

(1)Department of Chemical and Biological Engineering, Chalmers University of

Technology, Gothenburg 412 96, Sweden. nielsenj@chalmers.se.

BACKGROUND: The analysis of high-throughput data in biology is aided by

integrative approaches such as gene-set analysis. Gene-sets can represent

well-defined biological entities (e.g. metabolites) that interact in networks

(e.g. metabolic networks), to exert their function within the cell. Data

interpretation can benefit from incorporating the underlying network, but there

are currently no optimal methods that link gene-set analysis and network

structures.

RESULTS: Here we present Kiwi, a new tool that processes output data from

gene-set analysis and integrates them with a network structure such that the

inherent connectivity between gene-sets, i.e. not simply the gene overlap,

becomes apparent. In two case studies, we demonstrate that standard gene-set

analysis points at metabolites regulated in the interrogated condition.

Nevertheless, only the integration of the interactions between these metabolites

provides an extra layer of information that highlights how they are tightly

connected in the metabolic network.

CONCLUSIONS: Kiwi is a tool that enhances interpretability of high-throughput

data. It allows the users not only to discover a list of significant entities or

processes as in gene-set analysis, but also to visualize whether these entities

or processes are isolated or connected by means of their biological interaction.

Kiwi is available as a Python package at http://www.sysbio.se/kiwi and an online

tool in the BioMet Toolbox at http://www.biomet-toolbox.org.

DOI: 10.1186/s12859-014-0408-9

PMCID: PMC4269931

PMID: 25496126 [Indexed for MEDLINE]

685. Acta Crystallogr D Biol Crystallogr. 2014 Dec 1;70(Pt 12):3290-8. doi:

10.1107/S1399004714024572. Epub 2014 Nov 28.

ACHESYM: an algorithm and server for standardized placement of macromolecular

models in the unit cell.

Kowiel M(1), Jaskolski M(2), Dauter Z(3).

Author information:

(1)Department of Organic Chemistry, Poznan University of Medical Sciences,

Poznan, Poland. (2)Department of Crystallography, Faculty of Chemistry, A.

Mickiewicz University, Poznan, Poland. (3)Synchrotron Radiation Research Section,

Macromolecular Crystallography Laboratory, NCI, Argonne National Laboratory,

Argonne, USA.

Despite the existence of numerous useful conventions in structural

crystallography, for example for the choice of the asymmetric part of the unit

cell or of reciprocal space, surprisingly no standards are in use for the

placement of the molecular model in the unit cell, often leading to

inconsistencies or confusion. A conceptual solution for this problem has been

proposed for macromolecular crystal structures based on the idea of the

anti-Cheshire unit cell. Here, a program and server (called ACHESYM;

http://achesym.ibch.poznan.pl) are presented for the practical implementation of

this concept. In addition, the first task of ACHESYM is to find an optimal

(compact) macromolecular assembly if more than one polymer chain exists. ACHESYM

processes PDB (atomic parameters and TLS matrices) and mmCIF (diffraction data)

input files to produce a new coordinate set and to reindex the reflections and

modify their phases, if necessary.

DOI: 10.1107/S1399004714024572

PMCID: PMC4257622

PMID: 25478846 [Indexed for MEDLINE]

686. Anim Genet. 2014 Dec;45(6):898-902. doi: 10.1111/age.12208. Epub 2014 Sep 3.

Locus minimization in breed prediction using artificial neural network approach.

Iquebal MA(1), Ansari MS, Sarika, Dixit SP, Verma NK, Aggarwal RA, Jayakumar S,

Rai A, Kumar D.

Author information:

(1)Centre for Agricultural Bioinformatics, Indian Agricultural Statistics

Research Institute, Library Avenue, PUSA, New Delhi, 110012, India.

Molecular markers, viz. microsatellites and single nucleotide polymorphisms, have

revolutionized breed identification through the use of small samples of

biological tissue or germplasm, such as blood, carcass samples, embryos, ova and

semen, that show no evident phenotype. Classical tools of molecular data analysis

for breed identification have limitations, such as the unavailability of referral

breed data, causing increased cost of collection each time, compromised

computational accuracy and complexity of the methodology used. We report here the

successful use of an artificial neural network (ANN) in background to decrease

the cost of genotyping by locus minimization. The webserver is freely accessible

(http://nabg.iasri.res.in/bisgoat) to the research community. We demonstrate that

the machine learning (ANN) approach for breed identification is capable of

multifold advantages such as locus minimization, leading to a drastic reduction

in cost, and web availability of reference breed data, alleviating the need for

repeated genotyping each time one investigates the identity of an unknown breed.

To develop this model web implementation based on ANN, we used 51,850 samples of

allelic data of microsatellite-marker-based DNA fingerprinting on 25 loci

covering 22 registered goat breeds of India for training. Minimizing loci to up

to nine loci through the use of a multilayer perceptron model, we achieved 96.63%

training accuracy. This server can be an indispensable tool for identification of

existing breeds and new synthetic commercial breeds, leading to protection of

intellectual property in case of sovereignty and bio-piracy disputes. This server

can be widely used as a model for cost reduction by locus minimization for

various other flora and fauna in terms of variety, breed and/or line

identification, especially in conservation and improvement programs.

© 2014 Stichting International Foundation for Animal Genetics.

DOI: 10.1111/age.12208

PMID: 25183434 [Indexed for MEDLINE]

687. Bioinformatics. 2014 Dec 1;30(23):3410-1. doi: 10.1093/bioinformatics/btu572.

Epub 2014 Aug 24.

PHOXTRACK-a tool for interpreting comprehensive datasets of post-translational

modifications of proteins.

Weidner C(1), Fischer C(1), Sauer S(1).

Author information:

(1)Otto Warburg Laboratory, Max Planck Institute for Molecular Genetics,

Ihnestrasse 63-73, 14195 Berlin, Germany.

We introduce PHOXTRACK (PHOsphosite-X-TRacing Analysis of Causal Kinases), a

user-friendly freely available software tool for analyzing large datasets of

post-translational modifications of proteins, such as phosphorylation, which are

commonly gained by mass spectrometry detection. In contrast to other currently

applied data analysis approaches, PHOXTRACK uses full sets of quantitative

proteomics data and applies non-parametric statistics to calculate whether

defined kinase-specific sets of phosphosite sequences indicate statistically

significant concordant differences between various biological conditions.

PHOXTRACK is an efficient tool for extracting post-translational information of

comprehensive proteomics datasets to decipher key regulatory proteins and to

infer biologically relevant molecular pathways.AVAILABILITY: PHOXTRACK will be

maintained over the next years and is freely available as an online tool for

non-commercial use at http://phoxtrack.molgen.mpg.de. Users will also find a

tutorial at this Web site and can additionally give feedback at

https://groups.google.com/d/forum/phoxtrack-discuss.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu572

PMID: 25152232 [Indexed for MEDLINE]

688. Bioinformatics. 2014 Dec 1;30(23):3356-64. doi: 10.1093/bioinformatics/btu550.

Epub 2014 Aug 22.

SUBAcon: a consensus algorithm for unifying the subcellular localization data of

the Arabidopsis proteome.

Hooper CM(1), Tanz SK(2), Castleden IR(1), Vacher MA(2), Small ID(2), Millar

AH(1).

Author information:

(1)Centre of Excellence in Computational Systems Biology, The University of

Western Australia, Perth, WA 6009, Australia and ARC Centre of Excellence in

Plant Energy Biology, The University of Western Australia, Perth, WA 6009,

Australia. (2)Centre of Excellence in Computational Systems Biology, The

University of Western Australia, Perth, WA 6009, Australia and ARC Centre of

Excellence in Plant Energy Biology, The University of Western Australia, Perth,

WA 6009, Australia Centre of Excellence in Computational Systems Biology, The

University of Western Australia, Perth, WA 6009, Australia and ARC Centre of

Excellence in Plant Energy Biology, The University of Western Australia, Perth,

WA 6009, Australia.

MOTIVATION: Knowing the subcellular location of proteins is critical for

understanding their function and developing accurate networks representing

eukaryotic biological processes. Many computational tools have been developed to

predict proteome-wide subcellular location, and abundant experimental data from

green fluorescent protein (GFP) tagging or mass spectrometry (MS) are available

in the model plant, Arabidopsis. None of these approaches is error-free, and

thus, results are often contradictory.

RESULTS: To help unify these multiple data sources, we have developed the

SUBcellular Arabidopsis consensus (SUBAcon) algorithm, a naive Bayes classifier

that integrates 22 computational prediction algorithms, experimental GFP and MS

localizations, protein-protein interaction and co-expression data to derive a

consensus call and probability. SUBAcon classifies protein location in

Arabidopsis more accurately than single predictors.

AVAILABILITY: SUBAcon is a useful tool for recovering proteome-wide subcellular

locations of Arabidopsis proteins and is displayed in the SUBA3 database

(http://suba.plantenergy.uwa.edu.au). The source code and input data is available

through the SUBA3 server (http://suba.plantenergy.uwa.edu.au//SUBAcon.html) and

the Arabidopsis SUbproteome REference (ASURE) training set can be accessed using

the ASURE web portal (http://suba.plantenergy.uwa.edu.au/ASURE).

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu550

PMID: 25150248 [Indexed for MEDLINE]

689. Comput Biol Med. 2014 Dec;55:86-91. doi: 10.1016/j.compbiomed.2014.10.001. Epub

2014 Oct 14.

Haemophilus influenzae Genome Database (HIGDB): a single point web resource for

Haemophilus influenzae.

Swetha RG(1), Kala Sekar DK(2), Ramaiah S(1), Anbarasu A(3), Sekar K(2).

Author information:

(1)Medical & Biological Computing Laboratory, School of Biosciences and

Technology, VIT University, Vellore 632 014, India. (2)Laboratory for Structural

Biology and Bio-computing, Supercomputer Education and Research Centre, Indian

Institute of Science, Bangalore 560 012, India. (3)Medical & Biological Computing

Laboratory, School of Biosciences and Technology, VIT University, Vellore 632

014, India. Electronic address: aanand@vit.ac.in.

BACKGROUND: Haemophilus influenzae (H. Influenzae) is the causative agent of

pneumonia, bacteraemia and meningitis. The organism is responsible for large

number of deaths in both developed and developing countries. Even-though the

first bacterial genome to be sequenced was that of H. Influenzae, there is no

exclusive database dedicated for H. Influenzae. This prompted us to develop the

Haemophilus influenzae Genome Database (HIGDB).

METHODS: All data of HIGDB are stored and managed in MySQL database. The HIGDB is

hosted on Solaris server and developed using PERL modules. Ajax and JavaScript

are used for the interface development.

RESULTS: The HIGDB contains detailed information on 42,741 proteins, 18,077 genes

including 10 whole genome sequences and also 284 three dimensional structures of

proteins of H. influenzae. In addition, the database provides "Motif search" and

"GBrowse". The HIGDB is freely accessible through the URL:

http://bioserver1.physics.iisc.ernet.in/HIGDB/.

DISCUSSION: The HIGDB will be a single point access for bacteriological,

clinical, genomic and proteomic information of H. influenzae. The database can

also be used to identify DNA motifs within H. influenzae genomes and to compare

gene or protein sequences of a particular strain with other strains of H.

influenzae.

Copyright © 2014 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiomed.2014.10.001

PMID: 25450223 [Indexed for MEDLINE]

690. Dev Dyn. 2014 Dec;243(12):1632-6. doi: 10.1002/dvdy.24183. Epub 2014 Sep 30.

Poly peak parser: Method and software for identification of unknown indels using

sanger sequencing of polymerase chain reaction products.

Hill JT(1), Demarest BL, Bisgrove BW, Su YC, Smith M, Yost HJ.

Author information:

(1)Molecular Medicine Program and Department of Neurobiology & Anatomy,

University of Utah School of Medicine, Salt Lake City, Utah.

BACKGROUND: Genome editing techniques, including ZFN, TALEN, and CRISPR, have

created a need to rapidly screen many F1 individuals to identify carriers of

indels and determine the sequences of the mutations. Current techniques require

multiple clones of the targeted region to be sequenced for each individual, which

is inefficient when many individuals must be analyzed. Direct Sanger sequencing

of a polymerase chain reaction (PCR) amplified region surrounding the target site

is efficient, but Sanger sequencing genomes heterozygous for an indel results in

a string of "double peaks" due to the mismatched region.

RESULTS: To facilitate indel identification, we developed an online tool called

Poly Peak Parser (available at http://yost.genetics.utah.edu/software.php) that

is able to separate chromatogram data containing ambiguous base calls into

wild-type and mutant allele sequences. This tool allows the nature of the indel

to be determined from a single sequencing run per individual performed directly

on a PCR product spanning the targeted site, without cloning.

CONCLUSIONS: The method and algorithm described here facilitate rapid

identification and sequence characterization of heterozygous mutant carriers

generated by genome editing. Although designed for screening F1 individuals, this

tool can also be used to identify heterozygous indels in many contexts.

© 2014 The Authors. Developmental Dynamics published by Wiley Periodicals, Inc.

on behalf of American Association of Anatomists.

DOI: 10.1002/dvdy.24183

PMCID: PMC4525701

PMID: 25160973 [Indexed for MEDLINE]

691. Hum Mutat. 2014 Dec;35(12):1418-26. doi: 10.1002/humu.22693.

Mutation update: the spectra of nebulin variants and associated myopathies.

Lehtokari VL(1), Kiiski K, Sandaradura SA, Laporte J, Repo P, Frey JA, Donner K,

Marttila M, Saunders C, Barth PG, den Dunnen JT, Beggs AH, Clarke NF, North KN,

Laing NG, Romero NB, Winder TL, Pelin K, Wallgren-Pettersson C.

Author information:

(1)The Folkhälsan Institute of Genetics and the Department of Medical Genetics,

Haartman Institute, University of Helsinki, Helsinki, Finland.

A mutation update on the nebulin gene (NEB) is necessary because of recent

developments in analysis methodology, the identification of increasing numbers

and novel types of variants, and a widening in the spectrum of clinical and

histological phenotypes associated with this gigantic, 183 exons containing gene.

Recessive pathogenic variants in NEB are the major cause of nemaline myopathy

(NM), one of the most common congenital myopathies. Moreover, pathogenic NEB

variants have been identified in core-rod myopathy and in distal myopathies. In

this update, we present the disease-causing variants in NEB in 159 families, 143

families with NM, and 16 families with NM-related myopathies. Eighty-eight

families are presented here for the first time. We summarize 86 previously

published and 126 unpublished variants identified in NEB. Furthermore, we have

analyzed the NEB variants deposited in the Exome Variant Server

(http://evs.gs.washington.edu/EVS/), identifying that pathogenic variants are a

minor fraction of all coding variants (∼7%). This indicates that nebulin

tolerates substantial changes in its amino acid sequence, providing an

explanation as to why variants in such a large gene result in relatively rare

disorders. Lastly, we discuss the difficulties of drawing reliable

genotype-phenotype correlations in NEB-associated disease.

© 2014 WILEY PERIODICALS, INC.

DOI: 10.1002/humu.22693

PMCID: PMC4295925

PMID: 25205138 [Indexed for MEDLINE]

692. Mol Biosyst. 2014 Dec;10(12):3075-80. doi: 10.1039/c4mb00447g.

piRNA identification based on motif discovery.

Liu X(1), Ding J, Gong F.

Author information:

(1)School of Mathematics and Physics, University of Science and Technology

Beijing, Beijing 100083, China. mathlxq@163.com.

Piwi-interacting RNA (piRNA) is a class of small non-coding RNAs about 24 to 32

nucleotides long, associated with PIWI proteins, which are involved in germline

development, transposon silencing, and epigenetic regulation. Identification of

piRNA loci on the genome is very useful for further studies in the biogenesis and

function of piRNAs. To accomplish this, we applied the computational biology tool

Teiresias to identify motifs of variable length appearing frequently in mouse

piRNA and non-piRNA sequences, respectively, and then proposed an algorithm for

piRNA identification based on motif discovery, termed "Pibomd" by using these

sequence motifs as features in the Support Vector Machine (SVM) algorithm, a

sensitivity of 91.48% and a specificity of 89.76% on a mouse test dataset could

be achieved, much better results than those reported in previously published

algorithms. We also trained an unbalanced SVM classifier (named as "Asym-Pibomd")

that provided a higher specificity (96.2%) and a lower sensitivity (72.68%) than

Pibomd. Inspite of the predicted ACC being less than that of Pibomd, the

predicted ACC (84.44%) of Asym-Pibomd is about ten percent more than that

obtained using the k-mer method. Further analysis of the motif positions on the

piRNA sequences showed that the piRNA sequences may contain information at the

5'- and/or 3'-end recognized by the piRNA processing apparatus of actual piRNA

precursors. Furthermore, this prediction method can be found on a user-friendly

web server found at http://app.aporc.org/Pibomd/.

DOI: 10.1039/c4mb00447g

PMID: 25230731 [Indexed for MEDLINE]

693. Nat Chem Biol. 2014 Dec;10(12):1066-72. doi: 10.1038/nchembio.1666. Epub 2014 Oct

26.

Covalent docking of large libraries for the discovery of chemical probes.

London N(1), Miller RM(2), Krishnan S(3), Uchida K(3), Irwin JJ(4), Eidam O(1),

Gibold L(5), Cimermančič P(6), Bonnet R(7), Shoichet BK(4), Taunton J(3).

Author information:

(1)Department of Pharmaceutical Chemistry, University of California-San

Francisco, San Francisco, California, USA. (2)Chemistry and Chemical Biology

Graduate Program, University of California-San Francisco, San Francisco,

California, USA. (3)1] Department of Cellular and Molecular Pharmacology,

University of California-San Francisco, San Francisco, California, USA. [2]

Howard Hughes Medical Institute, University of California-San Francisco, San

Francisco, California, USA. (4)1] Department of Pharmaceutical Chemistry,

University of California-San Francisco, San Francisco, California, USA. [2]

Faculty of Pharmacy, University of Toronto, Toronto, Canada. [3] Ontario

Institute for Cancer Research, University of Toronto, Toronto, Canada. (5)1]

Faculty of Pharmacy, University of Toronto, Toronto, Canada. [2] Ontario

Institute for Cancer Research, University of Toronto, Toronto, Canada. [3]

Clermont Université, UMR 1071 INSERM-Université d'Auvergne, Clermont-Ferrand,

France. (6)Department of Cellular and Molecular Pharmacology, University of

California-San Francisco, San Francisco, California, USA. (7)1] Clermont

Université, UMR 1071 INSERM-Université d'Auvergne, Clermont-Ferrand, France. [2]

INRA, USC 2018, Clermont-Ferrand, France. [3] Service de Bactériologie, Centre

Hospitalier Universitaire, Clermont-Ferrand, France.

Erratum in

Nat Chem Biol. 2015 Mar;11(3):235.

Chemical probes that form a covalent bond with a protein target often show

enhanced selectivity, potency and utility for biological studies. Despite these

advantages, protein-reactive compounds are usually avoided in high-throughput

screening campaigns. Here we describe a general method (DOCKovalent) for

screening large virtual libraries of electrophilic small molecules. We apply this

method prospectively to discover reversible covalent fragments that target

distinct protein nucleophiles, including the catalytic serine of AmpC β-lactamase

and noncatalytic cysteines in RSK2, MSK1 and JAK3 kinases. We identify

submicromolar to low-nanomolar hits with high ligand efficiency, cellular

activity and selectivity, including what are to our knowledge the first reported

reversible covalent inhibitors of JAK3. Crystal structures of inhibitor complexes

with AmpC and RSK2 confirm the docking predictions and guide further

optimization. As covalent virtual screening may have broad utility for the rapid

discovery of chemical probes, we have made the method freely available through an

automated web server (http://covalent.docking.org/).

DOI: 10.1038/nchembio.1666

PMCID: PMC4232467

PMID: 25344815 [Indexed for MEDLINE]

694. Nucleic Acids Res. 2014 Dec 1;42(21):12961-72. doi: 10.1093/nar/gku1019. Epub

2014 Oct 31.

iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in

prokaryote with pseudo k-tuple nucleotide composition.

Lin H(1), Deng EZ(2), Ding H(2), Chen W(3), Chou KC(4).

Author information:

(1)Key Laboratory for Neuro-Information of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu 610054, China Gordon Life Science

Institute, Belmont, MA, USA hlin@gordonlifescience.org hlin@uestc.edu.cn. (2)Key

Laboratory for Neuro-Information of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu 610054, China. (3)Department of Physics,

School of Sciences, and Center for Genomics and Computational Biology, Hebei

United University, Tangshan 063000, China Gordon Life Science Institute, Belmont,

MA, USA hlin@gordonlifescience.org hlin@uestc.edu.cn. (4)Gordon Life Science

Institute, Belmont, MA, USA Center of Excellence in Genomic Medicine Research

(CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

hlin@gordonlifescience.org hlin@uestc.edu.cn.

The σ(54) promoters are unique in prokaryotic genome and responsible for

transcripting carbon and nitrogen-related genes. With the avalanche of genome

sequences generated in the postgenomic age, it is highly desired to develop

automated methods for rapidly and effectively identifying the σ(54) promoters.

Here, a predictor called 'iPro54-PseKNC' was developed. In the predictor, the

samples of DNA sequences were formulated by a novel feature vector called 'pseudo

k-tuple nucleotide composition', which was further optimized by the incremental

feature selection procedure. The performance of iPro54-PseKNC was examined by the

rigorous jackknife cross-validation tests on a stringent benchmark data set. As a

user-friendly web-server, iPro54-PseKNC is freely accessible at

http://lin.uestc.edu.cn/server/iPro54-PseKNC. For the convenience of the vast

majority of experimental scientists, a step-by-step protocol guide was provided

on how to use the web-server to get the desired results without the need to

follow the complicated mathematics that were presented in this paper just for its

integrity. Meanwhile, we also discovered through an in-depth statistical analysis

that the distribution of distances between the transcription start sites and the

translation initiation sites were governed by the gamma distribution, which may

provide a fundamental physical principle for studying the σ(54) promoters.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku1019

PMCID: PMC4245931

PMID: 25361964 [Indexed for MEDLINE]

695. Rev Sci Tech. 2014 Dec;33(3):937-46.

The equine arteritis virus isolate from the 2010 Argentinian outbreak.

Metz GE, Serena MS, Panei CJ, Nosetto EO, Echeverria MG.

A semen sample from a stallion infected during the 2010 equine arteritis virus

(EAV) outbreak was received for viral isolation prior to castration of the

animal. The virus was identified using a polyclonal antibody immunofluorescence

test. Reverse-transcription polymerase chain reaction (RT-PCR) was used to

amplify a region of the GP5 gene with primers GL105F and GL673R. The PCR products

were purified and sequences of both strands were determined in a MegaBACE™1000

with inner primers CR2 and EAV32. A phylogenetic dataset was built with the

previously reported sequences of five strains isolated in Argentina, together

with a group of selected sequences obtained from GenBank. The unrooted

neighbour-joining tree was constructed using molecular evolutionary genetic

analysis (MEGA) and bootstrap analyses were conducted using 1,000 replicate

datasets. Evolutionary distances were computed using the maximum composite

likelihood method. A NetNGlyc server analysis at the Technical University of

Denmark (www.cbs.dtu.dk/services/NetNGlyc/) was used to predict N-glycosylation

in GP5 sequences. The phylogenetic analysis revealed that the new strain

GLD-LP-ARG), together with other strains previously isolated, belongs to the

European group EU-1 but in a different branch. The new strain shows 99%

nucleotide identity with strain Al1and 98.1% with the Belgian strain 08P178.

Persistently infected stallions and their cryopreserved semen constitute a

reservoir of EAV, which ensures its persistence in the horse population around

the world. These findings reinforce the importance of careful monitoring of

persistently infected stallions, as well as semen straws, by RT-PCR or test

mating, in accordance with national regulations.

PMID: 25812217 [Indexed for MEDLINE]

696. BMC Bioinformatics. 2014 Nov 28;15:381. doi: 10.1186/s12859-014-0381-3.

CAR: contig assembly of prokaryotic draft genomes using rearrangements.

Lu CL(1), Chen KT(2), Huang SY(3), Chiu HT(4).

Author information:

(1)Department of Computer Science, National Tsing Hua University, Hsinchu, 300,

Taiwan. cllu@cs.nthu.edu.tw. (2)Department of Computer Science, National Tsing

Hua University, Hsinchu, 300, Taiwan. holystu@gmail.com. (3)Department of

Computer Science, National Tsing Hua University, Hsinchu, 300, Taiwan.

happyday5296@hotmail.com. (4)Department of Chemistry, National Cheng Kung

University, Tainan City, 701, Taiwan. hchiu@mail.ncku.edu.tw.

BACKGROUND: Next generation sequencing technology has allowed efficient

production of draft genomes for many organisms of interest. However, most draft

genomes are just collections of independent contigs, whose relative positions and

orientations along the genome being sequenced are unknown. Although several tools

have been developed to order and orient the contigs of draft genomes, more

accurate tools are still needed.

RESULTS: In this study, we present a novel reference-based contig assembly (or

scaffolding) tool, named as CAR, that can efficiently and more accurately order

and orient the contigs of a prokaryotic draft genome based on a reference genome

of a related organism. Given a set of contigs in multi-FASTA format and a

reference genome in FASTA format, CAR can output a list of scaffolds, each of

which is a set of ordered and oriented contigs. For validation, we have tested

CAR on a real dataset composed of several prokaryotic genomes and also compared

its performance with several other reference-based contig assembly tools.

Consequently, our experimental results have shown that CAR indeed performs better

than all these other reference-based contig assembly tools in terms of

sensitivity, precision and genome coverage.

CONCLUSIONS: CAR serves as an efficient tool that can more accurately order and

orient the contigs of a prokaryotic draft genome based on a reference genome. The

web server of CAR is freely available at http://genome.cs.nthu.edu.tw/CAR/ and

its stand-alone program can also be downloaded from the same website.

DOI: 10.1186/s12859-014-0381-3

PMCID: PMC4253983

PMID: 25431302 [Indexed for MEDLINE]

697. BMC Bioinformatics. 2014 Nov 25;15:362. doi: 10.1186/s12859-014-0362-6.

A statistical approach for 5' splice site prediction using short sequence motifs

and without encoding sequence data.

Meher PK(1), Sahu TK(2), Rao AR(3), Wahi SD(4).

Author information:

(1)Division of Statistical Genetics, Indian Agricultural Statistics Research

Institute, New Delhi, 110012, India. meherprabin@yahoo.com. (2)Centre for

Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute,

New Delhi, 110012, India. tanmayabioinfo@gmail.com. (3)Centre for Agricultural

Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi,

110012, India. rao.cshl.work@gmail.com. (4)Division of Statistical Genetics,

Indian Agricultural Statistics Research Institute, New Delhi, 110012, India.

sdwahi@iasri.res.in.

BACKGROUND: Most of the approaches for splice site prediction are based on

machine learning techniques. Though, these approaches provide high prediction

accuracy, the window lengths used are longer in size. Hence, these approaches may

not be suitable to predict the novel splice variants using the short sequence

reads generated from next generation sequencing technologies. Further, machine

learning techniques require numerically encoded data and produce different

accuracy with different encoding procedures. Therefore, splice site prediction

with short sequence motifs and without encoding sequence data became a motivation

for the present study.

RESULTS: An approach for finding association among nucleotide bases in the splice

site motifs is developed and used further to determine the appropriate window

size. Besides, an approach for prediction of donor splice sites using sum of

absolute error criterion has also been proposed. The proposed approach has been

compared with commonly used approaches i.e., Maximum Entropy Modeling (MEM),

Maximal Dependency Decomposition (MDD), Weighted Matrix Method (WMM) and Markov

Model of first order (MM1) and was found to perform equally with MEM and MDD and

better than WMM and MM1 in terms of prediction accuracy.

CONCLUSIONS: The proposed prediction approach can be used in the prediction of

donor splice sites with higher accuracy using short sequence motifs and hence can

be used as a complementary method to the existing approaches. Based on the

proposed methodology, a web server was also developed for easy prediction of

donor splice sites by users and is available at http://cabgrid.res.in:8080/sspred

.

DOI: 10.1186/s12859-014-0362-6

PMCID: PMC4702320

PMID: 25420551 [Indexed for MEDLINE]

698. PLoS One. 2014 Nov 24;9(11):e113811. doi: 10.1371/journal.pone.0113811.

eCollection 2014.

Protein loop modeling using a new hybrid energy function and its application to

modeling in inaccurate structural environments.

Park H(1), Lee GR(1), Heo L(1), Seok C(1).

Author information:

(1)Department of Chemistry, Seoul National University, Seoul, Republic of Korea.

Protein loop modeling is a tool for predicting protein local structures of

particular interest, providing opportunities for applications involving protein

structure prediction and de novo protein design. Until recently, the majority of

loop modeling methods have been developed and tested by reconstructing loops in

frameworks of experimentally resolved structures. In many practical applications,

however, the protein loops to be modeled are located in inaccurate structural

environments. These include loops in model structures, low-resolution

experimental structures, or experimental structures of different functional

forms. Accordingly, discrepancies in the accuracy of the structural environment

assumed in development of the method and that in practical applications present

additional challenges to modern loop modeling methods. This study demonstrates a

new strategy for employing a hybrid energy function combining physics-based and

knowledge-based components to help tackle this challenge. The hybrid energy

function is designed to combine the strengths of each energy component,

simultaneously maintaining accurate loop structure prediction in a

high-resolution framework structure and tolerating minor environmental errors in

low-resolution structures. A loop modeling method based on global optimization of

this new energy function is tested on loop targets situated in different levels

of environmental errors, ranging from experimental structures to structures

perturbed in backbone as well as side chains and template-based model structures.

The new method performs comparably to force field-based approaches in loop

reconstruction in crystal structures and better in loop prediction in inaccurate

framework structures. This result suggests that higher-accuracy predictions would

be possible for a broader range of applications. The web server for this method

is available at http://galaxy.seoklab.org/loop with the PS2 option for the

scoring function.

DOI: 10.1371/journal.pone.0113811

PMCID: PMC4242723

PMID: 25419655 [Indexed for MEDLINE]

699. BMC Bioinformatics. 2014 Nov 19;15:359. doi: 10.1186/s12859-014-0359-1.

Vindel: a simple pipeline for checking indel redundancy.

Li Z(1), Wu X(2), He B(3), Zhang L(4).

Author information:

(1)Department of Computer Science, Virginia Tech, Blacksburg, VA, 24061, USA.

zli04@vt.edu. (2)Department of Statistics, Virginia Tech, Blacksburg, VA, 24061,

USA. xwwu@vt.edu. (3)Department of Computer Science, Virginia Tech, Blacksburg,

VA, 24061, USA. binhe@vt.edu. (4)Department of Computer Science, Virginia Tech,

Blacksburg, VA, 24061, USA. lqzhang@vt.edu.

BACKGROUND: With the advance of next generation sequencing (NGS) technologies, a

large number of insertion and deletion (indel) variants have been identified in

human populations. Despite much research into variant calling, it has been found

that a non-negligible proportion of the identified indel variants might be false

positives due to sequencing errors, artifacts caused by ambiguous alignments, and

annotation errors.

RESULTS: In this paper, we examine indel redundancy in dbSNP, one of the central

databases for indel variants, and develop a standalone computational pipeline,

dubbed Vindel, to detect redundant indels. The pipeline first applies indel

position information to form candidate redundant groups, then performs indel

mutations to the reference genome to generate corresponding indel variant

substrings. Finally the indel variant substrings in the same candidate redundant

groups are compared in a pairwise fashion to identify redundant indels. We

applied our pipeline to check for redundancy in the human indels in dbSNP. Our

pipeline identified approximately 8% redundancy in insertion type indels, 12% in

deletion type indels, and overall 10% for insertions and deletions combined.

These numbers are largely consistent across all human autosomes. We also

investigated indel size distribution and adjacent indel distance distribution for

a better understanding of the mechanisms generating indel variants.

CONCLUSIONS: Vindel, a simple yet effective computational pipeline, can be used

to check whether a set of indels are redundant with respect to those already in

the database of interest such as NCBI's dbSNP. Of the approximately 5.9 million

indels we examined, nearly 0.6 million are redundant, revealing a serious

limitation in the current indel annotation. Statistics results prove the

consistency of the pipeline on indel redundancy detection for all 22 chromosomes.

Apart from the standalone Vindel pipeline, the indel redundancy check algorithm

is also implemented in the web server

http://bioinformatics.cs.vt.edu/zhanglab/indelRedundant.php .

DOI: 10.1186/s12859-014-0359-1

PMCID: PMC4245841

PMID: 25407965 [Indexed for MEDLINE]

700. BMC Bioinformatics. 2014 Nov 19;15:350. doi: 10.1186/s12859-014-0350-x.

Using phylogenetically-informed annotation (PIA) to search for light-interacting

genes in transcriptomes from non-model organisms.

Speiser DI(1,)(2), Pankey MS(3), Zaharoff AK(4), Battelle BA(5), Bracken-Grissom

HD(6), Breinholt JW(7), Bybee SM(8), Cronin TW(9), Garm A(10), Lindgren AR(11),

Patel NH(12), Porter ML(13), Protas ME(14), Rivera AS(15), Serb JM(16), Zigler

KS(17), Crandall KA(18,)(19), Oakley TH(20).

Author information:

(1)Department of Ecology, Evolution, and Marine Biology, University of California

Santa Barbara, Santa Barbara, CA, USA. speiser@mailbox.sc.edu. (2)Department of

Biological Sciences, University of South Carolina, Columbia, SC, USA.

speiser@mailbox.sc.edu. (3)Department of Ecology, Evolution, and Marine Biology,

University of California Santa Barbara, Santa Barbara, CA, USA.

abrinas@gmail.com. (4)Department of Ecology, Evolution, and Marine Biology,

University of California Santa Barbara, Santa Barbara, CA, USA.

sashazaharoff@gmail.com. (5)The Whitney Laboratory for Marine Bioscience,

University of Florida, St. Augustine, FL, USA. Battelle@whitney.ufl.edu.

(6)Department of Biological Sciences, Florida International University-Biscayne

Bay Campus, North Miami, FL, USA. hbracken@fiu.edu. (7)Florida Museum of Natural

History, University of Florida, Gainesville, FL, USA. jessebreinholt@gmail.com.

(8)Department of Biology, Brigham Young University, Provo, UT, USA.

seth.bybee@gmail.com. (9)Department of Biological Sciences, University of

Maryland Baltimore County, Baltimore, MD, USA. cronin@umbc.edu. (10)Department of

Biology, Marine Biological Section, University of Copenhagen, Copenhagen,

Denmark. algarm@bio.ku.dk. (11)Department of Biology, Portland State University,

Portland, OR, USA. arl3@pdx.edu. (12)Department of Molecular and Cell Biology &

Department of Integrative Biology, University of California, Berkeley, CA, USA.

nipam@uclink.berkeley.edu. (13)Department of Biology, University of South Dakota,

Vermillion, SD, USA. Megan.Porter@usd.edu. (14)Department of Natural Sciences and

Mathematics, Dominican University of California, San Rafael, CA, USA.

meredith.protas@dominican.edu. (15)Department of Biology, University of the

Pacific, Stockton, CA, USA. arivera@pacific.edu. (16)Department of Ecology,

Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA.

serb@iastate.edu. (17)Department of Biology, Sewanee: The University of the

South, Sewanee, TN, USA. kzigler@sewanee.edu. (18)Computational Biology

Institute, George Washington University, Ashburn, VA, USA. kcrandall@gwu.edu.

(19)Department of Invertebrate Zoology, National Museum of Natural History,

Smithsonian Institution, Washington, DC, USA. kcrandall@gwu.edu. (20)Department

of Ecology, Evolution, and Marine Biology, University of California Santa

Barbara, Santa Barbara, CA, USA. todd.oakley@lifesci.ucsb.edu.

BACKGROUND: Tools for high throughput sequencing and de novo assembly make the

analysis of transcriptomes (i.e. the suite of genes expressed in a tissue)

feasible for almost any organism. Yet a challenge for biologists is that it can

be difficult to assign identities to gene sequences, especially from non-model

organisms. Phylogenetic analyses are one useful method for assigning identities

to these sequences, but such methods tend to be time-consuming because of the

need to re-calculate trees for every gene of interest and each time a new data

set is analyzed. In response, we employed existing tools for phylogenetic

analysis to produce a computationally efficient, tree-based approach for

annotating transcriptomes or new genomes that we term Phylogenetically-Informed

Annotation (PIA), which places uncharacterized genes into pre-calculated

phylogenies of gene families.

RESULTS: We generated maximum likelihood trees for 109 genes from a Light

Interaction Toolkit (LIT), a collection of genes that underlie the function or

development of light-interacting structures in metazoans. To do so, we searched

protein sequences predicted from 29 fully-sequenced genomes and built trees using

tools for phylogenetic analysis in the Osiris package of Galaxy (an open-source

workflow management system). Next, to rapidly annotate transcriptomes from

organisms that lack sequenced genomes, we repurposed a maximum likelihood-based

Evolutionary Placement Algorithm (implemented in RAxML) to place sequences of

potential LIT genes on to our pre-calculated gene trees. Finally, we implemented

PIA in Galaxy and used it to search for LIT genes in 28 newly-sequenced

transcriptomes from the light-interacting tissues of a range of cephalopod

mollusks, arthropods, and cubozoan cnidarians. Our new trees for LIT genes are

available on the Bitbucket public repository (

http://bitbucket.org/osiris\_phylogenetics/pia/ ) and we demonstrate PIA on a

publicly-accessible web server ( http://galaxy-dev.cnsi.ucsb.edu/pia/ ).

CONCLUSIONS: Our new trees for LIT genes will be a valuable resource for

researchers studying the evolution of eyes or other light-interacting structures.

We also introduce PIA, a high throughput method for using phylogenetic

relationships to identify LIT genes in transcriptomes from non-model organisms.

With simple modifications, our methods may be used to search for different sets

of genes or to annotate data sets from taxa outside of Metazoa.

DOI: 10.1186/s12859-014-0350-x

PMCID: PMC4255452

PMID: 25407802 [Indexed for MEDLINE]

701. Database (Oxford). 2014 Nov 19;2014:bau104. doi: 10.1093/database/bau104. Print

2014.

Kin-Driver: a database of driver mutations in protein kinases.

Simonetti FL(1), Tornador C(1), Nabau-Moretó N(1), Molina-Vila MA(1),

Marino-Buslje C(2).

Author information:

(1)Fundación Instituto Leloir, Av. Patricias Argentinas 435. C1405BWE, Buenos

Aires, Argentina, Pompeu Fabra University (UPF), Dept. de Tecnologies de la

Informació i les Comunicacions. Tanger 122-140 08018, Barcelona, Spain,

Computational Genomics Laboratory, Genetics Department, Institut de Biologia

Universitat de Barcelona (IBUB), Facultat de Biologia, Av Diagonal 645 and

Breakthrough Cancer Research Unit, Dexeus University Hospital, Sabino Arana 5-19,

Barcelona, Spain. (2)Fundación Instituto Leloir, Av. Patricias Argentinas 435.

C1405BWE, Buenos Aires, Argentina, Pompeu Fabra University (UPF), Dept. de

Tecnologies de la Informació i les Comunicacions. Tanger 122-140 08018,

Barcelona, Spain, Computational Genomics Laboratory, Genetics Department,

Institut de Biologia Universitat de Barcelona (IBUB), Facultat de Biologia, Av

Diagonal 645 and Breakthrough Cancer Research Unit, Dexeus University Hospital,

Sabino Arana 5-19, Barcelona, Spain. cmb@leloir.org.ar.

Somatic mutations in protein kinases (PKs) are frequent driver events in many

human tumors, while germ-line mutations are associated with hereditary diseases.

Here we present Kin-driver, the first database that compiles driver mutations in

PKs with experimental evidence demonstrating their functional role. Kin-driver is

a manual expert-curated database that pays special attention to activating

mutations (AMs) and can serve as a validation set to develop new generation tools

focused on the prediction of gain-of-function driver mutations. It also offers an

easy and intuitive environment to facilitate the visualization and analysis of

mutations in PKs. Because all mutations are mapped onto a multiple sequence

alignment, analogue positions between kinases can be identified and tentative new

mutations can be proposed for studying by transferring annotation. Finally, our

database can also be of use to clinical and translational laboratories, helping

them to identify uncommon AMs that can correlate with response to new antitumor

drugs. The website was developed using PHP and JavaScript, which are supported by

all major browsers; the database was built using MySQL server. Kin-driver is

available at: http://kin-driver.leloir.org.ar/

© The Author(s) 2014. Published by Oxford University Press.

DOI: 10.1093/database/bau104

PMCID: PMC4237945

PMID: 25414382 [Indexed for MEDLINE]

702. Genome Biol Evol. 2014 Nov 18;6(12):3199-209. doi: 10.1093/gbe/evu252.

Indel reliability in indel-based phylogenetic inference.

Ashkenazy H(1), Cohen O(2), Pupko T(3), Huchon D(4).

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, Ramat Aviv, Israel. (2)Department of Cell Research

and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University,

Ramat Aviv, Israel Present address: Department of Molecular Genetics, Weizmann

Institute of Science, Rehovot, Israel. (3)Department of Cell Research and

Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat

Aviv, Israel talp@post.tau.ac.il huchond@post.tau.ac.il. (4)Department of

Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat

Aviv, Israel talp@post.tau.ac.il huchond@post.tau.ac.il.

It is often assumed that it is unlikely that the same insertion or deletion

(indel) event occurred at the same position in two independent evolutionary

lineages, and thus, indel-based inference of phylogeny should be less subject to

homoplasy compared with standard inference which is based on substitution events.

Indeed, indels were successfully used to solve debated evolutionary relationships

among various taxonomical groups. However, indels are never directly observed but

rather inferred from the alignment and thus indel-based inference may be

sensitive to alignment errors. It is hypothesized that phylogenetic

reconstruction would be more accurate if it relied only on a subset of reliable

indels instead of the entire indel data. Here, we developed a method to quantify

the reliability of indel characters by measuring how often they appear in a set

of alternative multiple sequence alignments. Our approach is based on the

assumption that indels that are consistently present in most alternative

alignments are more reliable compared with indels that appear only in a small

subset of these alignments. Using simulated and empirical data, we studied the

impact of filtering and weighting indels by their reliability scores on the

accuracy of indel-based phylogenetic reconstruction. The new method is available

as a web-server at http://guidance.tau.ac.il/RELINDEL/.

© The Author(s) 2014. Published by Oxford University Press on behalf of the

Society for Molecular Biology and Evolution.

DOI: 10.1093/gbe/evu252

PMCID: PMC4986452

PMID: 25409663 [Indexed for MEDLINE]

703. Bioinformatics. 2014 Nov 15;30(22):3215-22. doi: 10.1093/bioinformatics/btu508.

Epub 2014 Jul 30.

Walking the interactome for candidate prioritization in exome sequencing studies

of Mendelian diseases.

Smedley D(1), Köhler S(1), Czeschik JC(1), Amberger J(1), Bocchini C(1), Hamosh

A(1), Veldboer J(2), Zemojtel T(2), Robinson PN(3).

Author information:

(1)Mouse Informatics Group, The Wellcome Trust Sanger Institute, Wellcome Trust

Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK, Institute for Medical

Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger

Platz 1, 13353 Berlin, Genome Informatics Department, Institute of Human

Genetics, University Hospital Essen, University of Duisburg-Essen, Hufelandstr.

55, 45122 Essen, Germany, McKusick-Nathans Institute of Genetic Medicine, John

Hopkins University School of Medicine, Baltimore, MD 21205, USA, Department of

Mathematics and Computer Science, Institute for Bioinformatics, Freie Universität

Berlin, Takustrasse 9, 14195 Berlin, Germany, Institute of Bioorganic Chemistry,

Polish Academy of Sciences, 61-701 Poznan, Poland, Berlin-Brandenburg Center for

Regenerative Therapies, Charité-Universitätsmedizin Berlin, Augustenburger Platz

1, 13353 Berlin and Max Planck Institute for Molecular Genetics, Ihnestrasse 73,

14195 Berlin, Germany. (2)Mouse Informatics Group, The Wellcome Trust Sanger

Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK,

Institute for Medical Genetics and Human Genetics, Charité-Universitätsmedizin

Berlin, Augustenburger Platz 1, 13353 Berlin, Genome Informatics Department,

Institute of Human Genetics, University Hospital Essen, University of

Duisburg-Essen, Hufelandstr. 55, 45122 Essen, Germany, McKusick-Nathans Institute

of Genetic Medicine, John Hopkins University School of Medicine, Baltimore, MD

21205, USA, Department of Mathematics and Computer Science, Institute for

Bioinformatics, Freie Universität Berlin, Takustrasse 9, 14195 Berlin, Germany,

Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-701 Poznan,

Poland, Berlin-Brandenburg Center for Regenerative Therapies,

Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin and Max

Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Mouse Informatics Group, The Wellcome Trust Sanger Institute, Wellcome Trust

Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK, Institute for Medical

Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger

Platz 1, 13353 Berlin, Genome Informatics Department, Institute of Human

Genetics, University Hospital Essen, University of Duisburg-Essen, Hufelandstr.

55, 45122 Essen, Germany, McKusick-Nathans Institute of Genetic Medicine, John

Hopkins University School of Medicine, Baltimore, MD 21205, USA, Department of

Mathematics and Computer Science, Institute for Bioinformatics, Freie Universität

Berlin, Takustrasse 9, 14195 Berlin, Germany, Institute of Bioorganic Chemistry,

Polish Academy of Sciences, 61-701 Poznan, Poland, Berlin-Brandenburg Center for

Regenerative Therapies, Charité-Universitätsmedizin Berlin, Augustenburger Platz

1, 13353 Berlin and Max Planck Institute for Molecular Genetics, Ihnestrasse 73,

14195 Berlin, Germany. (3)Mouse Informatics Group, The Wellcome Trust Sanger

Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK,

Institute for Medical Genetics and Human Genetics, Charité-Universitätsmedizin

Berlin, Augustenburger Platz 1, 13353 Berlin, Genome Informatics Department,

Institute of Human Genetics, University Hospital Essen, University of

Duisburg-Essen, Hufelandstr. 55, 45122 Essen, Germany, McKusick-Nathans Institute

of Genetic Medicine, John Hopkins University School of Medicine, Baltimore, MD

21205, USA, Department of Mathematics and Computer Science, Institute for

Bioinformatics, Freie Universität Berlin, Takustrasse 9, 14195 Berlin, Germany,

Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-701 Poznan,

Poland, Berlin-Brandenburg Center for Regenerative Therapies,

Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin and Max

Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Mouse Informatics Group, The Wellcome Trust Sanger Institute, Wellcome Trust

Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK, Institute for Medical

Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger

Platz 1, 13353 Berlin, Genome Informatics Department, Institute of Human

Genetics, University Hospital Essen, University of Duisburg-Essen, Hufelandstr.

55, 45122 Essen, Germany, McKusick-Nathans Institute of Genetic Medicine, John

Hopkins University School of Medicine, Baltimore, MD 21205, USA, Department of

Mathematics and Computer Science, Institute for Bioinformatics, Freie Universität

Berlin, Takustrasse 9, 14195 Berlin, Germany, Institute of Bioorganic Chemistry,

Polish Academy of Sciences, 61-701 Poznan, Poland, Berlin-Brandenburg Center for

Regenerative Therapies, Charité-Universitätsmedizin Berlin, Augustenburger Platz

1, 13353 Berlin and Max Planck Institute for Molecular Genetics, Ihnestrasse 73,

14195 Berlin, Germany Mouse Informatics Group, The Wellcome Trust Sang

MOTIVATION: Whole-exome sequencing (WES) has opened up previously unheard of

possibilities for identifying novel disease genes in Mendelian disorders, only

about half of which have been elucidated to date. However, interpretation of WES

data remains challenging.

RESULTS: Here, we analyze protein-protein association (PPA) networks to identify

candidate genes in the vicinity of genes previously implicated in a disease. The

analysis, using a random-walk with restart (RWR) method, is adapted to the

setting of WES by developing a composite variant-gene relevance score based on

the rarity, location and predicted pathogenicity of variants and the RWR

evaluation of genes harboring the variants. Benchmarking using known disease

variants from 88 disease-gene families reveals that the correct gene is ranked

among the top 10 candidates in ≥50% of cases, a figure which we confirmed using a

prospective study of disease genes identified in 2012 and PPA data produced

before that date. We implement our method in a freely available Web server,

ExomeWalker, that displays a ranked list of candidates together with information

on PPAs, frequency and predicted pathogenicity of the variants to allow quick and

effective searches for candidates that are likely to reward closer investigation.

AVAILABILITY AND IMPLEMENTATION: http://compbio.charite.de/ExomeWalker

CONTACT: : peter.robinson@charite.de.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu508

PMCID: PMC4221119

PMID: 25078397 [Indexed for MEDLINE]

704. Database (Oxford). 2014 Nov 11;2014. pii: bau105. doi: 10.1093/database/bau105.

Print 2014.

iPathCons and iPathDB: an improved insect pathway construction tool and the

database.

Zhang Z(1), Yin C(1), Liu Y(1), Jie W(1), Lei W(1), Li F(2).

Author information:

(1)Department of Entomology, College of Plant Protection, Nanjing Agricultural

University and The Key laboratory of Monitoring and Management of Plant Diseases

and Insects, Ministry of Agriculture, No. 1, Weigang Road, Xuanwu District,

Nanjing, Jiangsu 210095, China. (2)Department of Entomology, College of Plant

Protection, Nanjing Agricultural University and The Key laboratory of Monitoring

and Management of Plant Diseases and Insects, Ministry of Agriculture, No. 1,

Weigang Road, Xuanwu District, Nanjing, Jiangsu 210095, China lifei@njau.edu.cn.

Insects are one of the most successful animal groups on earth. Some insects, such

as the silkworm and honeybee, are beneficial to humans, whereas others are

notorious pests of crops. At present, the genomes of 38 insects have been

sequenced and made publically available. In addition, the transcriptomes of

dozens of insects have been sequenced. As gene data rapidly accumulate,

constructing the pathway of molecular interactions becomes increasingly important

for entomological research. Here, we developed an improved tool, iPathCons, for

knowledge-based construction of pathways from the transcriptomes or the official

gene sets of genomes. Considering the high evolution diversity in insects,

iPathCons uses a voting system for Kyoto Encyclopedia of Genes and Genomes

Orthology assignment. Both stand-alone software and a web server of iPathCons are

provided. Using iPathCons, we constructed the pathways of molecular interactions

of 52 insects, including 37 genome-sequenced and 15 transcriptome-sequenced ones.

These pathways are available in the iPathDB, which provides searches, web server,

data downloads, etc. This database will be highly useful for the insect research

community. Database URL: http://ento.njau.edu.cn/ipath/

© The Author(s) 2014. Published by Oxford University Press.

DOI: 10.1093/database/bau105

PMCID: PMC4227299

PMID: 25388589 [Indexed for MEDLINE]

705. J Biotechnol. 2014 Nov 10;189:154-6. doi: 10.1016/j.jbiotec.2014.09.012. Epub

2014 Sep 28.

CEM-designer: design of custom expression microarrays in the post-ENCODE Era.

Arnold C(1), Externbrink F(2), Hackermüller J(3), Reiche K(4).

Author information:

(1)Bioinformatics Group, Department for Computer Science, University of Leipzig,

Leipzig, Germany; RNomics Group, Department of Diagnostics, Fraunhofer Institute

for Cell Therapy and Immunology - IZI, Leipzig, Germany. (2)Bioinformatics Group,

Department for Computer Science, University of Leipzig, Leipzig, Germany.

(3)Young Investigators Group, Bioinformatics and Transcriptomics, Department

Proteomics, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany;

Bioinformatics Group, Department for Computer Science, University of Leipzig,

Leipzig, Germany; RNomics Group, Department of Diagnostics, Fraunhofer Institute

for Cell Therapy and Immunology - IZI, Leipzig, Germany. Electronic address:

joerg.hackermueller@ufz.de. (4)Young Investigators Group, Bioinformatics and

Transcriptomics, Department Proteomics, Helmholtz Centre for Environmental

Research - UFZ, Leipzig, Germany; Bioinformatics Group, Department for Computer

Science, University of Leipzig, Leipzig, Germany; RNomics Group, Department of

Diagnostics, Fraunhofer Institute for Cell Therapy and Immunology - IZI, Leipzig,

Germany. Electronic address: kristin.reiche@ufz.de.

Microarrays are widely used in gene expression studies, and custom expression

microarrays are popular to monitor expression changes of a customer-defined set

of genes. However, the complexity of transcriptomes uncovered recently make

custom expression microarray design a non-trivial task. Pervasive transcription

and alternative processing of transcripts generate a wealth of interweaved

transcripts that requires well-considered probe design strategies and is largely

neglected in existing approaches. We developed the web server CEM-Designer that

facilitates microarray platform independent design of custom expression

microarrays for complex transcriptomes. CEM-Designer covers (i) the collection

and generation of a set of unique target sequences from different sources and

(ii) the selection of a set of sensitive and specific probes that optimally

represents the target sequences. Probe design itself is left to third party

software to ensure that probes meet provider-specific constraints. CEM-Designer

is available at http://designpipeline.bioinf.uni-leipzig.de.

Copyright © 2014 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.jbiotec.2014.09.012

PMID: 25262644 [Indexed for MEDLINE]

706. BMC Bioinformatics. 2014 Nov 8;15:354. doi: 10.1186/s12859-014-0354-6.

MLGO: phylogeny reconstruction and ancestral inference from gene-order data.

Hu F(1,)(2), Lin Y(3), Tang J(4,)(5).

Author information:

(1)Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin

University, Tianjin, 300072, China. hufeiyc@gmail.com. (2)Department of Computer

Science and Engineering, University of South Carolina, Columbia, 29208, SC, USA.

hufeiyc@gmail.com. (3)Department of Computer Science and Engineering, University

of California, San Diego, 92093 La Jolla, CA, USA. yul280@ucsd.edu. (4)Tianjin

Key Laboratory of Cognitive Computing and Application, Tianjin University,

Tianjin, 300072, China. jtang@ces.sc.edu. (5)Department of Computer Science and

Engineering, University of South Carolina, Columbia, 29208, SC, USA.

jtang@ces.sc.edu.

BACKGROUND: The rapid accumulation of whole-genome data has renewed interest in

the study of using gene-order data for phylogenetic analyses and ancestral

reconstruction. Current software and web servers typically do not support

duplication and loss events along with rearrangements.

RESULTS: MLGO (Maximum Likelihood for Gene-Order Analysis) is a web tool for the

reconstruction of phylogeny and/or ancestral genomes from gene-order data. MLGO

is based on likelihood computation and shows advantages over existing methods in

terms of accuracy, scalability and flexibility.

CONCLUSIONS: To the best of our knowledge, it is the first web tool for analysis

of large-scale genomic changes including not only rearrangements but also gene

insertions, deletions and duplications. The web tool is available from

http://www.geneorder.org/server.php .

DOI: 10.1186/s12859-014-0354-6

PMCID: PMC4236499

PMID: 25376663 [Indexed for MEDLINE]

707. Database (Oxford). 2014 Nov 7;2014. pii: bau108. doi: 10.1093/database/bau108.

Print 2014.

The Virtual Xenbase: transitioning an online bioinformatics resource to a private

cloud.

Karimi K(1), Vize PD(2).

Author information:

(1)Department of Biological Sciences, University of Calgary, 2500 University Dr

NW, Calgary, Alberta, Canada, T2N 1N4. (2)Department of Biological Sciences,

University of Calgary, 2500 University Dr NW, Calgary, Alberta, Canada, T2N 1N4

kkarimi@ucalgary.ca pvize@ucalgary.ca.

As a model organism database, Xenbase has been providing informatics and genomic

data on Xenopus (Silurana) tropicalis and Xenopus laevis frogs for more than a

decade. The Xenbase database contains curated, as well as community-contributed

and automatically harvested literature, gene and genomic data. A GBrowse genome

browser, a BLAST+ server and stock center support are available on the site. When

this resource was first built, all software services and components in Xenbase

ran on a single physical server, with inherent reliability, scalability and

inter-dependence issues. Recent advances in networking and virtualization

techniques allowed us to move Xenbase to a virtual environment, and more

specifically to a private cloud. To do so we decoupled the different software

services and components, such that each would run on a different virtual machine.

In the process, we also upgraded many of the components. The resulting system is

faster and more reliable. System maintenance is easier, as individual virtual

machines can now be updated, backed up and changed independently. We are also

experiencing more effective resource allocation and utilization. Database URL:

www.xenbase.org.

© The Author(s) 2014. Published by Oxford University Press.

DOI: 10.1093/database/bau108

PMCID: PMC4224262

PMID: 25380782 [Indexed for MEDLINE]

708. Database (Oxford). 2014 Nov 7;2014. pii: bau107. doi: 10.1093/database/bau107.

Print 2014.

ChloroSSRdb: a repository of perfect and imperfect chloroplastic simple sequence

repeats (cpSSRs) of green plants.

Kapil A(1), Rai PK(1), Shanker A(2).

Author information:

(1)Department of Bioinformatics, Banasthali University, Banasthali, Rajasthan

304022, India and Department of Mathematics and Statistics, Banasthali

University, Banasthali, Rajasthan 304022, India. (2)Department of Bioinformatics,

Banasthali University, Banasthali, Rajasthan 304022, India and Department of

Mathematics and Statistics, Banasthali University, Banasthali, Rajasthan 304022,

India ashomics@gmail.com.

Simple sequence repeats (SSRs) are regions in DNA sequence that contain repeating

motifs of length 1-6 nucleotides. These repeats are ubiquitously present and are

found in both coding and non-coding regions of genome. A total of 534 complete

chloroplast genome sequences (as on 18 September 2014) of Viridiplantae are

available at NCBI organelle genome resource. It provides opportunity to mine

these genomes for the detection of SSRs and store them in the form of a database.

In an attempt to properly manage and retrieve chloroplastic SSRs, we designed

ChloroSSRdb which is a relational database developed using SQL server 2008 and

accessed through ASP.NET. It provides information of all the three types

(perfect, imperfect and compound) of SSRs. At present, ChloroSSRdb contains

124 430 mined SSRs, with majority lying in non-coding region. Out of these, PCR

primers were designed for 118 249 SSRs. Tetranucleotide repeats (47 079) were

found to be the most frequent repeat type, whereas hexanucleotide repeats (6414)

being the least abundant. Additionally, in each species statistical analyses were

performed to calculate relative frequency, correlation coefficient and chi-square

statistics of perfect and imperfect SSRs. In accordance with the growing interest

in SSR studies, ChloroSSRdb will prove to be a useful resource in developing

genetic markers, phylogenetic analysis, genetic mapping, etc. Moreover, it will

serve as a ready reference for mined SSRs in available chloroplast genomes of

green plants. Database URL: www.compubio.in/chlorossrdb/

© The Author(s) 2014. Published by Oxford University Press.

DOI: 10.1093/database/bau107

PMCID: PMC4224265

PMID: 25380781 [Indexed for MEDLINE]

709. J Theor Biol. 2014 Nov 7;360:34-45. doi: 10.1016/j.jtbi.2014.06.031. Epub 2014

Jul 2.

R3P-Loc: a compact multi-label predictor using ridge regression and random

projection for protein subcellular localization.

Wan S(1), Mak MW(2), Kung SY(3).

Author information:

(1)Department of Electronic and Information Engineering, The Hong Kong

Polytechnic University, Hong Kong SAR, China. Electronic address:

10900600r@connect.polyu.hk. (2)Department of Electronic and Information

Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China.

Electronic address: enmwmak@polyu.edu.hk. (3)Department of Electrical

Engineering, Princeton University, NJ, USA. Electronic address:

kung@princeton.edu.

Locating proteins within cellular contexts is of paramount significance in

elucidating their biological functions. Computational methods based on knowledge

databases (such as gene ontology annotation (GOA) database) are known to be more

efficient than sequence-based methods. However, the predominant scenarios of

knowledge-based methods are that (1) knowledge databases typically have enormous

size and are growing exponentially, (2) knowledge databases contain redundant

information, and (3) the number of extracted features from knowledge databases is

much larger than the number of data samples with ground-truth labels. These

properties render the extracted features liable to redundant or irrelevant

information, causing the prediction systems suffer from overfitting. To address

these problems, this paper proposes an efficient multi-label predictor, namely

R3P-Loc, which uses two compact databases for feature extraction and applies

random projection (RP) to reduce the feature dimensions of an ensemble ridge

regression (RR) classifier. Two new compact databases are created from Swiss-Prot

and GOA databases. These databases possess almost the same amount of information

as their full-size counterparts but with much smaller size. Experimental results

on two recent datasets (eukaryote and plant) suggest that R3P-Loc can reduce the

dimensions by seven-folds and significantly outperforms state-of-the-art

predictors. This paper also demonstrates that the compact databases reduce the

memory consumption by 39 times without causing degradation in prediction

accuracy. For readers׳ convenience, the R3P-Loc server is available online at

url:http://bioinfo.eie.polyu.edu.hk/R3PLocServer/.

Copyright © 2014 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2014.06.031

PMID: 24997236 [Indexed for MEDLINE]

710. J Biomol NMR. 2014 Nov;60(2-3):131-46. doi: 10.1007/s10858-014-9863-x. Epub 2014

Oct 2.

CSI 2.0: a significantly improved version of the Chemical Shift Index.

Hafsa NE(1), Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, Canada.

Protein chemical shifts have long been used by NMR spectroscopists to assist with

secondary structure assignment and to provide useful distance and torsion angle

constraint data for structure determination. One of the most widely used methods

for secondary structure identification is called the Chemical Shift Index (CSI).

The CSI method uses a simple digital chemical shift filter to locate secondary

structures along the protein chain using backbone (13)C and (1)H chemical shifts.

While the CSI method is simple to use and easy to implement, it is only about

75-80% accurate. Here we describe a significantly improved version of the CSI

(2.0) that uses machine-learning techniques to combine all six backbone chemical

shifts ((13)Cα, (13)Cβ, (13)C, (15)N, (1)HN, (1)Hα) with sequence-derived

features to perform far more accurate secondary structure identification. Our

tests indicate that CSI 2.0 achieved an average identification accuracy (Q3) of

90.56% for a training set of 181 proteins in a repeated tenfold cross-validation

and 89.35% for a test set of 59 proteins. This represents a significant

improvement over other state-of-the-art chemical shift-based methods. In

particular, the level of performance of CSI 2.0 is equal to that of standard

methods, such as DSSP and STRIDE, used to identify secondary structures via 3D

coordinate data. This suggests that CSI 2.0 could be used both in providing

accurate NMR constraint data in the early stages of protein structure

determination as well as in defining secondary structure locations in the final

protein model(s). A CSI 2.0 web server (http://csi.wishartlab.com) is available

for submitting the input queries for secondary structure identification.

DOI: 10.1007/s10858-014-9863-x

PMID: 25273503 [Indexed for MEDLINE]

711. J Biomol NMR. 2014 Nov;60(2-3):73-5. doi: 10.1007/s10858-014-9855-x. Epub 2014

Sep 5.

PONDEROSA-C/S: client-server based software package for automated protein 3D

structure determination.

Lee W(1), Stark JL, Markley JL.

Author information:

(1)National Magnetic Resonance Facility at Madison, and Biochemistry Department,

University of Wisconsin-Madison, Madison, WI, 53706, USA, whlee@nmrfam.wisc.edu.

Peak-picking Of Noe Data Enabled by Restriction Of Shift Assignments-Client

Server (PONDEROSA-C/S) builds on the original PONDEROSA software (Lee et al. in

Bioinformatics 27:1727-1728. doi: 10.1093/bioinformatics/btr200, 2011) and

includes improved features for structure calculation and refinement.

PONDEROSA-C/S consists of three programs: Ponderosa Server, Ponderosa Client, and

Ponderosa Analyzer. PONDEROSA-C/S takes as input the protein sequence, a list of

assigned chemical shifts, and nuclear Overhauser data sets ((13)C- and/or

(15)N-NOESY). The output is a set of assigned NOEs and 3D structural models for

the protein. Ponderosa Analyzer supports the visualization, validation, and

refinement of the results from Ponderosa Server. These tools enable

semi-automated NMR-based structure determination of proteins in a rapid and

robust fashion. We present examples showing the use of PONDEROSA-C/S in solving

structures of four proteins: two that enable comparison with the original

PONDEROSA package, and two from the Critical Assessment of automated Structure

Determination by NMR (Rosato et al. in Nat Methods 6:625-626. doi:

10.1038/nmeth0909-625 , 2009) competition. The software package can be downloaded

freely in binary format from http://pine.nmrfam.wisc.edu/download\_packages.html.

Registered users of the National Magnetic Resonance Facility at Madison can

submit jobs to the PONDEROSA-C/S server at http://ponderosa.nmrfam.wisc.edu,

where instructions, tutorials, and instructions can be found. Structures are

normally returned within 1-2 days.

DOI: 10.1007/s10858-014-9855-x

PMCID: PMC4207954

PMID: 25190042 [Indexed for MEDLINE]

712. Mitochondrion. 2014 Nov;19 Pt B:334-7. doi: 10.1016/j.mito.2014.02.002. Epub 2014

Feb 19.

MitoSatPlant: mitochondrial microsatellites database of viridiplantae.

Kumar M(1), Kapil A(2), Shanker A(3).

Author information:

(1)Department of Computer Science, Banasthali University, Banasthali, 304022

Rajasthan, India. (2)Department of Bioinformatics, Banasthali University,

Banasthali, 304022 Rajasthan, India. (3)Department of Bioinformatics, Banasthali

University, Banasthali, 304022 Rajasthan, India. Electronic address:

ashomics@gmail.com.

Microsatellites also known as simple sequence repeats (SSRs) consist of 1-6

nucleotide long repeating units. The importance of mitochondrial SSRs (mtSSRs) in

fields like population genetics, plant phylogenetics and genome mapping motivated

us to develop MitoSatPlant, a repository of plant mtSSRs. It contains information

for perfect, imperfect and compound SSRs mined from 92 mitochondrial genomes of

green plants, available at NCBI (as of 1 Feb 2014). A total of 72,798 SSRs were

found, of which PCR primers were designed for 72,495 SSRs. Among all sequences,

tetranucleotide repeats (26,802) were found to be most abundant whereas

hexanucleotide repeats (2751) were detected with least frequency. MitoSatPlant

was developed using SQL server 2008 and can be accessed through a front end

designed in ASP.Net. It is an easy to use, user-friendly database and will prove

to be a useful resource for plant scientists. To the best of our knowledge

MitoSatPlant is the only database available for plant mtSSRs and can be freely

accessed at http://compubio.in/mitosatplant/.

Copyright © 2014 Elsevier B.V. and Mitochondria Research Society. All rights

reserved.

DOI: 10.1016/j.mito.2014.02.002

PMID: 24561221 [Indexed for MEDLINE]

713. RNA. 2014 Nov;20(11):1666-70. doi: 10.1261/rna.043687.113. Epub 2014 Sep 18.

circBase: a database for circular RNAs.

Glažar P(1), Papavasileiou P(1), Rajewsky N(2).

Author information:

(1)Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany. (2)Max

Delbrück Center for Molecular Medicine, 13125 Berlin, Germany

rajewsky@mdc-berlin.de.

Recently, several laboratories have reported thousands of circular RNAs

(circRNAs) in animals. Numerous circRNAs are highly stable and have specific

spatiotemporal expression patterns. Even though a function for circRNAs is

unknown, these features make circRNAs an interesting class of RNAs as possible

biomarkers and for further research. We developed a database and website,

"circBase," where merged and unified data sets of circRNAs and the evidence

supporting their expression can be accessed, downloaded, and browsed within the

genomic context. circBase also provides scripts to identify known and novel

circRNAs in sequencing data. The database is freely accessible through the web

server at http://www.circbase.org/.

© 2014 Glažar et al.; Published by Cold Spring Harbor Laboratory Press for the

RNA Society.

DOI: 10.1261/rna.043687.113

PMCID: PMC4201819

PMID: 25234927 [Indexed for MEDLINE]

714. J Comput Chem. 2014 Oct 30;35(28):2040-6. doi: 10.1002/jcc.23718. Epub 2014 Sep

12.

Predicting backbone Cα angles and dihedrals from protein sequences by stacked

sparse auto-encoder deep neural network.

Lyons J(1), Dehzangi A, Heffernan R, Sharma A, Paliwal K, Sattar A, Zhou Y, Yang

Y.

Author information:

(1)Institute for Integrated and Intelligent Systems, Griffith University,

Brisbane, Australia.

Because a nearly constant distance between two neighbouring Cα atoms, local

backbone structure of proteins can be represented accurately by the angle between

C(αi-1)-C(αi)-C(αi+1) (θ) and a dihedral angle rotated about the C(αi)-C(αi+1)

bond (τ). θ and τ angles, as the representative of structural properties of three

to four amino-acid residues, offer a description of backbone conformations that

is complementary to φ and ψ angles (single residue) and secondary structures (>3

residues). Here, we report the first machine-learning technique for

sequence-based prediction of θ and τ angles. Predicted angles based on an

independent test have a mean absolute error of 9° for θ and 34° for τ with a

distribution on the θ-τ plane close to that of native values. The average

root-mean-square distance of 10-residue fragment structures constructed from

predicted θ and τ angles is only 1.9Å from their corresponding native structures.

Predicted θ and τ angles are expected to be complementary to predicted ϕ and ψ

angles and secondary structures for using in model validation and template-based

as well as template-free structure prediction. The deep neural network learning

technique is available as an on-line server called Structural Property prediction

with Integrated DEep neuRal network (SPIDER) at http://sparks-lab.org.

Copyright © 2014 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23718

PMID: 25212657 [Indexed for MEDLINE]

715. Nurs Manag (Harrow). 2014 Oct 30;21(7):15. doi: 10.7748/nm.21.7.15.s22.

Sickness absence.

[No authors listed]

NHS Employers has launched a free online tool to improve the management of staff

sick leave in the NHS. The tool is designed to help managers adopt a confident

and consistent approach, and provides step-by-step information about what to do

when staff call in sick, practical advice on some of the common reasons for

sickness absence, and information on what to do if staff are off sick frequently

or are on long-term sick leave. It also has information about managing return to

work and how to prevent sickness absence. Visit the microsite at

www.nhsemployers.org/sickness.

DOI: 10.7748/nm.21.7.15.s22

PMID: 25355120

716. BMC Genomics. 2014 Oct 23;15:925. doi: 10.1186/1471-2164-15-925.

SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences.

Agostini F, Cirillo D, Ponti RD, Tartaglia GG(1).

Author information:

(1)Gene Function and Evolution, Centre for Genomic Regulation (CRG), C/ Dr,

Aiguader 88, 08003 Barcelona, Spain. gian.tartaglia@crg.es.

BACKGROUND: The large amount of data produced by high-throughput sequencing poses

new computational challenges. In the last decade, several tools have been

developed for the identification of transcription and splicing factor binding

sites.

RESULTS: Here, we introduce the SeAMotE (Sequence Analysis of Motifs Enrichment)

algorithm for discovery of regulatory regions in nucleic acid sequences. SeAMotE

provides (i) a robust analysis of high-throughput sequence sets, (ii) a motif

search based on pattern occurrences and (iii) an easy-to-use web-server

interface. We applied our method to recently published data including 351

chromatin immunoprecipitation (ChIP) and 13 crosslinking immunoprecipitation

(CLIP) experiments and compared our results with those of other well-established

motif discovery tools. SeAMotE shows an average accuracy of 80% in finding

discriminative motifs and outperforms other methods available in literature.

CONCLUSIONS: SeAMotE is a fast, accurate and flexible algorithm for the

identification of sequence patterns involved in protein-DNA and protein-RNA

recognition. The server can be freely accessed at

http://s.tartaglialab.com/new\_submission/seamote.

DOI: 10.1186/1471-2164-15-925

PMCID: PMC4223730

PMID: 25341390 [Indexed for MEDLINE]

717. Bioinformatics. 2014 Oct 15;30(20):2989-90. doi: 10.1093/bioinformatics/btu428.

Epub 2014 Jul 4.

TAPAS: tools to assist the targeted protein quantification of human alternative

splice variants.

Yang JS(1), Sabidó E(1), Serrano L(2), Kiel C(1).

Author information:

(1)EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG),

Universitat Pompeu Fabra (UPF), Proteomics Unit, Centre for Genomic Regulation

(CRG), 08003 Barcelona and Institució Catalana de Recerca i Estudis Avançats

(ICREA), 08010 Barcelona, Spain EMBL/CRG Systems Biology Research Unit, Centre

for Genomic Regulation (CRG), Universitat Pompeu Fabra (UPF), Proteomics Unit,

Centre for Genomic Regulation (CRG), 08003 Barcelona and Institució Catalana de

Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain. (2)EMBL/CRG Systems

Biology Research Unit, Centre for Genomic Regulation (CRG), Universitat Pompeu

Fabra (UPF), Proteomics Unit, Centre for Genomic Regulation (CRG), 08003

Barcelona and Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010

Barcelona, Spain EMBL/CRG Systems Biology Research Unit, Centre for Genomic

Regulation (CRG), Universitat Pompeu Fabra (UPF), Proteomics Unit, Centre for

Genomic Regulation (CRG), 08003 Barcelona and Institució Catalana de Recerca i

Estudis Avançats (ICREA), 08010 Barcelona, Spain EMBL/CRG Systems Biology

Research Unit, Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra

(UPF), Proteomics Unit, Centre for Genomic Regulation (CRG), 08003 Barcelona and

Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona,

Spain.

MOTIVATION: In proteomes of higher eukaryotes, many alternative splice variants

can only be detected by their shared peptides. This makes it highly challenging

to use peptide-centric mass spectrometry to distinguish and to quantify protein

isoforms resulting from alternative splicing events.

RESULTS: We have developed two complementary algorithms based on linear

mathematical models to efficiently compute a minimal set of shared and unique

peptides needed to quantify a set of isoforms and splice variants. Further, we

developed a statistical method to estimate the splice variant abundances based on

stable isotope labeled peptide quantities. The algorithms and databases are

integrated in a web-based tool, and we have experimentally tested the limits of

our quantification method using spiked proteins and cell extracts.

AVAILABILITY AND IMPLEMENTATION: The TAPAS server is available at URL

http://davinci.crg.es/tapas/.

CONTACT: luis.serrano@crg.eu or christina.kiel@crg.eu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu428

PMID: 24996896 [Indexed for MEDLINE]

718. Bioinformatics. 2014 Oct 15;30(20):2973-4. doi: 10.1093/bioinformatics/btu411.

Epub 2014 Jun 27.

TPpred2: improving the prediction of mitochondrial targeting peptide cleavage

sites by exploiting sequence motifs.

Savojardo C(1), Martelli PL(1), Fariselli P(2), Casadio R(1).

Author information:

(1)Biocomputing Group, University of Bologna, CIRI-Health Science and

Technology/Department of Biology, 40126 Bologna and Department of Computer

Science and Engineering, University of Bologna, 40127 Bologna, Italy.

(2)Biocomputing Group, University of Bologna, CIRI-Health Science and

Technology/Department of Biology, 40126 Bologna and Department of Computer

Science and Engineering, University of Bologna, 40127 Bologna, Italy Biocomputing

Group, University of Bologna, CIRI-Health Science and Technology/Department of

Biology, 40126 Bologna and Department of Computer Science and Engineering,

University of Bologna, 40127 Bologna, Italy.

SUMMARY: Targeting peptides are N-terminal sorting signals in proteins that

promote their translocation to mitochondria through the interaction with

different protein machineries. We recently developed TPpred, a machine

learning-based method scoring among the best ones available to predict the

presence of a targeting peptide into a protein sequence and its cleavage site.

Here we introduce TPpred2 that improves TPpred performances in the task of

identifying the cleavage site of the targeting peptides. TPpred2 is now available

as a web interface and as a stand-alone version for users who can freely download

and adopt it for processing large volumes of sequences. Availability and

implementaion: TPpred2 is available both as web server and stand-alone version at

http://tppred2.biocomp.unibo.it.

CONTACT: gigi@biocomp.unibo.it

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu411

PMID: 24974200 [Indexed for MEDLINE]

719. BMC Bioinformatics. 2014 Oct 4;15:343. doi: 10.1186/1471-2105-15-343.

CLAP: a web-server for automatic classification of proteins with special

reference to multi-domain proteins.

Gnanavel M, Mehrotra P, Rakshambikai R, Martin J, Srinivasan N(1), Bhaskara RM.

Author information:

(1)Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012,

India. ns@mbu.iisc.ernet.in.

BACKGROUND: The function of a protein can be deciphered with higher accuracy from

its structure than from its amino acid sequence. Due to the huge gap in the

available protein sequence and structural space, tools that can generate

functionally homogeneous clusters using only the sequence information, hold great

importance. For this, traditional alignment-based tools work well in most cases

and clustering is performed on the basis of sequence similarity. But, in the case

of multi-domain proteins, the alignment quality might be poor due to varied

lengths of the proteins, domain shuffling or circular permutations. Multi-domain

proteins are ubiquitous in nature, hence alignment-free tools, which overcome the

shortcomings of alignment-based protein comparison methods, are required.

Further, existing tools classify proteins using only domain-level information and

hence miss out on the information encoded in the tethered regions or accessory

domains. Our method, on the other hand, takes into account the full-length

sequence of a protein, consolidating the complete sequence information to

understand a given protein better.

RESULTS: Our web-server, CLAP (Classification of Proteins), is one such

alignment-free software for automatic classification of protein sequences. It

utilizes a pattern-matching algorithm that assigns local matching scores (LMS) to

residues that are a part of the matched patterns between two sequences being

compared. CLAP works on full-length sequences and does not require prior domain

definitions.Pilot studies undertaken previously on protein kinases and

immunoglobulins have shown that CLAP yields clusters, which have high functional

and domain architectural similarity. Moreover, parsing at a statistically

determined cut-off resulted in clusters that corroborated with the sub-family

level classification of that particular domain family.

CONCLUSIONS: CLAP is a useful protein-clustering tool, independent of domain

assignment, domain order, sequence length and domain diversity. Our method can be

used for any set of protein sequences, yielding functionally relevant clusters

with high domain architectural homogeneity. The CLAP web server is freely

available for academic use at http://nslab.mbu.iisc.ernet.in/clap/.

DOI: 10.1186/1471-2105-15-343

PMCID: PMC4287353

PMID: 25282152 [Indexed for MEDLINE]

720. Anal Biochem. 2014 Oct 1;462:76-83. doi: 10.1016/j.ab.2014.06.022. Epub 2014 Jul

10.

iTIS-PseTNC: a sequence-based predictor for identifying translation initiation

site in human genes using pseudo trinucleotide composition.

Chen W(1), Feng PM(2), Deng EZ(3), Lin H(4), Chou KC(5).

Author information:

(1)Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, Hebei United University, Tangshan 063000, China; Gordon

Life Science Institute, Boston, MA 02478, USA. Electronic address:

chenweiimu@gmail.com. (2)School of Public Health, Hebei United University,

Tangshan 063000, China. Electronic address: fengpengmian@gmail.com. (3)Key

Laboratory for Neuro-Information of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu 610054, China. Electronic address:

enzeas@gmail.com. (4)Gordon Life Science Institute, Boston, MA 02478, USA; Key

Laboratory for Neuro-Information of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu 610054, China. Electronic address:

hlin@uestc.edu.cn. (5)Department of Physics, School of Sciences, Center for

Genomics and Computational Biology, Hebei United University, Tangshan 063000,

China; Gordon Life Science Institute, Boston, MA 02478, USA; Center of Excellence

in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589,

Saudi Arabia. Electronic address: kcchou@gordonlifescience.org.

Translation is a key process for gene expression. Timely identification of the

translation initiation site (TIS) is very important for conducting in-depth

genome analysis. With the avalanche of genome sequences generated in the

postgenomic age, it is highly desirable to develop automated methods for rapidly

and effectively identifying TIS. Although some computational methods were

proposed in this regard, none of them considered the global or long-range

sequence-order effects of DNA, and hence their prediction quality was limited. To

count this kind of effects, a new predictor, called "iTIS-PseTNC," was developed

by incorporating the physicochemical properties into the pseudo trinucleotide

composition, quite similar to the PseAAC (pseudo amino acid composition) approach

widely used in computational proteomics. It was observed by the rigorous

cross-validation test on the benchmark dataset that the overall success rate

achieved by the new predictor in identifying TIS locations was over 97%. As a web

server, iTIS-PseTNC is freely accessible at

http://lin.uestc.edu.cn/server/iTIS-PseTNC. To maximize the convenience of the

vast majority of experimental scientists, a step-by-step guide is provided on how

to use the web server to obtain the desired results without the need to go

through detailed mathematical equations, which are presented in this paper just

for the integrity of the new prection method.

Copyright © 2014 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2014.06.022

PMID: 25016190 [Indexed for MEDLINE]

721. Comput Biol Med. 2014 Oct;53:164-70. doi: 10.1016/j.compbiomed.2014.07.016. Epub

2014 Jul 31.

LRRsearch: An asynchronous server-based application for the prediction of

leucine-rich repeat motifs and an integrative database of NOD-like receptors.

Bej A(1), Sahoo BR(1), Swain B(1), Basu M(1), Jayasankar P(2), Samanta M(3).

Author information:

(1)Fish Health Management Division, Central Institute of Freshwater Aquaculture

(CIFA), Kausalyaganga, Bhubaneswar, Odisha 751002, India. (2)Central Institute of

Freshwater Aquaculture (CIFA), Kausalyaganga, Bhubaneswar, Odisha 751002, India.

(3)Fish Health Management Division, Central Institute of Freshwater Aquaculture

(CIFA), Kausalyaganga, Bhubaneswar, Odisha 751002, India. Electronic address:

msamanta1969@yahoo.com.

The leucine-rich repeat (LRR) motifs of the nucleotide-binding oligomerization

domain like receptors (NLRs) play key roles in recognizing and binding various

pathogen associated molecular patterns (PAMPs) resulting in the activation of

downstream signaling and innate immunity. Therefore, identification of LRR motifs

is very important to study ligand-receptor interaction. To date, available

resources pose restrictions including both false negative and false positive

prediction of LRR motifs from the primary protein sequence as their algorithms

are relied either only on sequence based comparison or alignment techniques or

are over biased for a particular LRR containing protein family. Therefore, to

minimize the error (≤5%) and to identify a maximum number of LRR motifs in the

wide range of proteins, we have developed "LRRsearch" web-server using position

specific scoring matrix (PSSM) of 11 residue LRR-HCS (highly conserved segment)

which are frequently observed motifs in the most divergent classes of LRR

containing proteins. A data library of 421 proteins, distributed among five known

NLR families has also been integrated with the "LRRsearch" for the rich user

experience. The access to the "LRRsearch" program is freely available at

http://www.lrrsearch.com/.

Copyright © 2014 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiomed.2014.07.016

PMID: 25150822 [Indexed for MEDLINE]

722. Genomics Proteomics Bioinformatics. 2014 Oct;12(5):249-53. doi:

10.1016/j.gpb.2014.08.001. Epub 2014 Sep 16.

nuMap: a web platform for accurate prediction of nucleosome positioning.

Alharbi BA(1), Alshammari TH(2), Felton NL(3), Zhurkin VB(4), Cui F(5).

Author information:

(1)Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology,

Rochester, NY 14623, USA. (2)B. Thomas Golisano College of Computing &

Information Sciences, Rochester Institute of Technology, Rochester, NY 14623,

USA. (3)Information Technology Services, Rochester Institute of Technology,

Rochester, NY 14623, USA. (4)Laboratory of Cell Biology, National Cancer

Institute, National Institutes of Health, Bethesda, MD 20892, USA. (5)Thomas H.

Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY

14623, USA. Electronic address: fxcsbi@rit.edu.

Nucleosome positioning is critical for gene expression and of major biological

interest. The high cost of experimentally mapping nucleosomal arrangement

signifies the need for computational approaches to predict nucleosome positions

at high resolution. Here, we present a web-based application to fulfill this need

by implementing two models, YR and W/S schemes, for the translational and

rotational positioning of nucleosomes, respectively. Our methods are based on

sequence-dependent anisotropic bending that dictates how DNA is wrapped around a

histone octamer. This application allows users to specify a number of options

such as schemes and parameters for threading calculation and provides multiple

layout formats. The nuMap is implemented in Java/Perl/MySQL and is freely

available for public use at http://numap.rit.edu. The user manual, implementation

notes, description of the methodology and examples are available at the site.

Copyright © 2014 The Authors. Production and hosting by Elsevier Ltd.. All rights

reserved.

DOI: 10.1016/j.gpb.2014.08.001

PMCID: PMC4411418

PMID: 25220945 [Indexed for MEDLINE]

723. J Digit Imaging. 2014 Oct;27(5):563-70. doi: 10.1007/s10278-014-9692-1.

A reliable, low-cost picture archiving and communications system for small and

medium veterinary practices built using open-source technology.

Iotti B(1), Valazza A.

Author information:

(1)Dipartimento di Scienze Veterinarie, Università degli Studi di Torino, Torino,

Italy, bryan.iotti@unito.it.

Picture Archiving and Communications Systems (PACS) are the most needed system in

a modern hospital. As an integral part of the Digital Imaging and Communications

in Medicine (DICOM) standard, they are charged with the responsibility for secure

storage and accessibility of the diagnostic imaging data. These machines need to

offer high performance, stability, and security while proving reliable and

ergonomic in the day-to-day and long-term storage and retrieval of the data they

safeguard. This paper reports the experience of the authors in developing and

installing a compact and low-cost solution based on open-source technologies in

the Veterinary Teaching Hospital for the University of Torino, Italy, during the

course of the summer of 2012. The PACS server was built on low-cost x86-based

hardware and uses an open source operating system derived from Oracle OpenSolaris

(Oracle Corporation, Redwood City, CA, USA) to host the DCM4CHEE PACS DICOM

server (DCM4CHEE, http://www.dcm4che.org ). This solution features very high data

security and an ergonomic interface to provide easy access to a large amount of

imaging data. The system has been in active use for almost 2 years now and has

proven to be a scalable, cost-effective solution for practices ranging from small

to very large, where the use of different hardware combinations allows scaling to

the different deployments, while the use of paravirtualization allows increased

security and easy migrations and upgrades.

DOI: 10.1007/s10278-014-9692-1

PMCID: PMC4171423

PMID: 24793019 [Indexed for MEDLINE]

724. Proteins. 2014 Oct;82(10):2565-73. doi: 10.1002/prot.24620. Epub 2014 Jun 19.

Direct prediction of profiles of sequences compatible with a protein structure by

neural networks with fragment-based local and energy-based nonlocal profiles.

Li Z(1), Yang Y, Faraggi E, Zhan J, Zhou Y.

Author information:

(1)School of Informatics and Computing, Indiana University-Purdue University,

Indianapolis, Indiana, 46202.

Locating sequences compatible with a protein structural fold is the well-known

inverse protein-folding problem. While significant progress has been made, the

success rate of protein design remains low. As a result, a library of designed

sequences or profile of sequences is currently employed for guiding experimental

screening or directed evolution. Sequence profiles can be computationally

predicted by iterative mutations of a random sequence to produce energy-optimized

sequences, or by combining sequences of structurally similar fragments in a

template library. The latter approach is computationally more efficient but

yields less accurate profiles than the former because of lacking tertiary

structural information. Here we present a method called SPIN that predicts

Sequence Profiles by Integrated Neural network based on fragment-derived sequence

profiles and structure-derived energy profiles. SPIN improves over the

fragment-derived profile by 6.7% (from 23.6 to 30.3%) in sequence identity

between predicted and wild-type sequences. The method also reduces the number of

residues in low complex regions by 15.7% and has a significantly better balance

of hydrophilic and hydrophobic residues at protein surface. The accuracy of

sequence profiles obtained is comparable to those generated from the protein

design program RosettaDesign 3.5. This highly efficient method for predicting

sequence profiles from structures will be useful as a single-body scoring term

for improving scoring functions used in protein design and fold recognition. It

also complements protein design programs in guiding experimental design of the

sequence library for screening and directed evolution of designed sequences. The

SPIN server is available at http://sparks-lab.org.

© 2014 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24620

PMCID: PMC4177274

PMID: 24898915 [Indexed for MEDLINE]

725. Proteins. 2014 Oct;82(10):2455-71. doi: 10.1002/prot.24610. Epub 2014 Jun 9.

RBRDetector: improved prediction of binding residues on RNA-binding protein

structures using complementary feature- and template-based strategies.

Yang XX(1), Deng ZL, Liu R.

Author information:

(1)Agricultural Bioinformatics Key Laboratory of Hubei Province, College of

Informatics, Huazhong Agricultural University, Wuhan, 430070, People's Republic

of China.

Computational prediction of RNA-binding residues is helpful in uncovering the

mechanisms underlying protein-RNA interactions. Traditional algorithms

individually applied feature- or template-based prediction strategy to recognize

these crucial residues, which could restrict their predictive power. To improve

RNA-binding residue prediction, herein we propose the first integrative algorithm

termed RBRDetector (RNA-Binding Residue Detector) by combining these two

strategies. We developed a feature-based approach that is an ensemble learning

predictor comprising multiple structure-based classifiers, in which well-defined

evolutionary and structural features in conjunction with sequential or structural

microenvironment were used as the inputs of support vector machines. Meanwhile,

we constructed a template-based predictor to recognize the putative RNA-binding

regions by structurally aligning the query protein to the RNA-binding proteins

with known structures. The final RBRDetector algorithm is an ingenious fusion of

our feature- and template-based approaches based on a piecewise function. By

validating our predictors with diverse types of structural data, including bound

and unbound structures, native and simulated structures, and protein structures

binding to different RNA functional groups, we consistently demonstrated that

RBRDetector not only had clear advantages over its component methods, but also

significantly outperformed the current state-of-the-art algorithms. Nevertheless,

the major limitation of our algorithm is that it performed relatively well on

DNA-binding proteins and thus incorrectly predicted the DNA-binding regions as

RNA-binding interfaces. Finally, we implemented the RBRDetector algorithm as a

user-friendly web server, which is freely accessible at

http://ibi.hzau.edu.cn/rbrdetector.

© 2014 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24610

PMID: 24854765 [Indexed for MEDLINE]

726. Bioinformation. 2014 Sep 30;10(9):602-5. doi: 10.6026/97320630010602. eCollection

2014.

TargetCompare: A web interface to compare simultaneous miRNAs targets.

Moreira FC(1), Dustan B(1), Hamoy IG(2), Ribeiro-Dos-Santos AM(3), Dos Santos

AR(4).

Author information:

(1)Laboratorio de Genética Humana e Medica, Instituto de Ciências Biológicas,

Universidade Federal do Para, Belem, PA, Brasil ; Centro Universitário do Pará -

Área de Ciências Exatas e Tecnologia, , Belém, PA, Brasil. (2)Laboratorio de

Genética Humana e Medica, Instituto de Ciências Biológicas, Universidade Federal

do Para, Belem, PA, Brasil ; Universidade Federal Rural da Amazônia, Campus de

Capanema, PA, Brasil. (3)Laboratorio de Genética Humana e Medica, Instituto de

Ciências Biológicas, Universidade Federal do Para, Belem, PA, Brasil.

(4)Laboratorio de Genética Humana e Medica, Instituto de Ciências Biológicas,

Universidade Federal do Para, Belem, PA, Brasil ; Nucleo de Pesquisa em

Oncologia, Universidade Federal do Para, Belem, PA, Brasil.

MicroRNAs (miRNAs) are small non-coding nucleotide sequences between 17 and 25

nucleotides in length that primarily function in the regulation of gene

expression. A since miRNA has thousand of predict targets in a complex,

regulatory cell signaling network. Therefore, it is of interest to study multiple

target genes simultaneously. Hence, we describe a web tool (developed using Java

programming language and MySQL database server) to analyse multiple targets of

pre-selected miRNAs. We cross validated the tool in eight most highly expressed

miRNAs in the antrum region of stomach. This helped to identify 43 potential

genes that are target of at least six of the referred miRNAs. The developed tool

aims to reduce the randomness and increase the chance of selecting strong

candidate target genes and miRNAs responsible for playing important roles in the

studied tissue.AVAILABILITY: http://lghm.ufpa.br/targetcompare.

DOI: 10.6026/97320630010602

PMCID: PMC4209372

PMID: 25352731

727. Front Genet. 2014 Sep 18;5:325. doi: 10.3389/fgene.2014.00325. eCollection 2014.

DaVIE: Database for the Visualization and Integration of Epigenetic data.

Fejes AP(1), Jones MJ(1), Kobor MS(2).

Author information:

(1)Centre for Molecular Medicine and Therapeutics, Child and Family Research

Institute, University of British Columbia Vancouver, BC, Canada. (2)Centre for

Molecular Medicine and Therapeutics, Child and Family Research Institute,

University of British Columbia Vancouver, BC, Canada ; Department of Medical

Genetics, University of British Columbia Vancouver, BC, Canada.

One of the challenges in the analysis of large data sets, particularly in a

population-based setting, is the ability to perform comparisons across projects.

This has to be done in such a way that the integrity of each individual project

is maintained, while ensuring that the data are comparable across projects. These

issues are beginning to be observed in human DNA methylation studies, as the

Illumina 450k platform and next generation sequencing-based assays grow in

popularity and decrease in price. This increase in productivity is enabling new

insights into epigenetics, but also requires the development of pipelines and

software capable of handling the large volumes of data. The specific problems

inherent in creating a platform for the storage, comparison, integration, and

visualization of DNA methylation data include data storage, algorithm efficiency

and ability to interpret the results to derive biological meaning from them.

Databases provide a ready-made solution to these issues, but as yet no tools

exist that that leverage these advantages while providing an intuitive user

interface for interpreting results in a genomic context. We have addressed this

void by integrating a database to store DNA methylation data with a web interface

to query and visualize the database and a set of libraries for more complex

analysis. The resulting platform is called DaVIE: Database for the Visualization

and Integration of Epigenetics data. DaVIE can use data culled from a variety of

sources, and the web interface includes the ability to group samples by sub-type,

compare multiple projects and visualize genomic features in relation to sites of

interest. We have used DaVIE to identify patterns of DNA methylation in specific

projects and across different projects, identify outlier samples, and cross-check

differentially methylated CpG sites identified in specific projects across large

numbers of samples. A demonstration server has been setup using GEO data at

http://echelon.cmmt.ubc.ca/dbaccess/, with login "guest" and password "guest."

Groups may download and install their own version of the server following the

instructions on the project's wiki.

DOI: 10.3389/fgene.2014.00325

PMCID: PMC4166999

PMID: 25278960

728. PLoS Comput Biol. 2014 Sep 18;10(9):e1003829. doi: 10.1371/journal.pcbi.1003829.

eCollection 2014.

eMatchSite: sequence order-independent structure alignments of ligand binding

pockets in protein models.

Brylinski M(1).

Author information:

(1)Department of Biological Sciences, Louisiana State University, Baton Rouge,

Louisiana, United States of America; Center for Computation & Technology,

Louisiana State University, Baton Rouge, Louisiana, United States of America.

Detecting similarities between ligand binding sites in the absence of global

homology between target proteins has been recognized as one of the critical

components of modern drug discovery. Local binding site alignments can be

constructed using sequence order-independent techniques, however, to achieve a

high accuracy, many current algorithms for binding site comparison require

high-quality experimental protein structures, preferably in the bound

conformational state. This, in turn, complicates proteome scale applications,

where only various quality structure models are available for the majority of

gene products. To improve the state-of-the-art, we developed eMatchSite, a new

method for constructing sequence order-independent alignments of ligand binding

sites in protein models. Large-scale benchmarking calculations using

adenine-binding pockets in crystal structures demonstrate that eMatchSite

generates accurate alignments for almost three times more protein pairs than

SOIPPA. More importantly, eMatchSite offers a high tolerance to structural

distortions in ligand binding regions in protein models. For example, the

percentage of correctly aligned pairs of adenine-binding sites in weakly

homologous protein models is only 4-9% lower than those aligned using crystal

structures. This represents a significant improvement over other algorithms, e.g.

the performance of eMatchSite in recognizing similar binding sites is 6% and 13%

higher than that of SiteEngine using high- and moderate-quality protein models,

respectively. Constructing biologically correct alignments using predicted ligand

binding sites in protein models opens up the possibility to investigate

drug-protein interaction networks for complete proteomes with prospective

systems-level applications in polypharmacology and rational drug repositioning.

eMatchSite is freely available to the academic community as a web-server and a

stand-alone software distribution at http://www.brylinski.org/ematchsite.

DOI: 10.1371/journal.pcbi.1003829

PMCID: PMC4168975

PMID: 25232727 [Indexed for MEDLINE]

729. PLoS One. 2014 Sep 18;9(9):e107889. doi: 10.1371/journal.pone.0107889.

eCollection 2014.

BioSWR--semantic web services registry for bioinformatics.

Repchevsky D(1), Gelpi JL(2).

Author information:

(1)Barcelona Supercomputing Center, Life-Sciences Department, National Institute

of Bioinformatics, Computational Bioinformatics Node, Barcelona, Spain.

(2)Barcelona Supercomputing Center, Life-Sciences Department, National Institute

of Bioinformatics, Computational Bioinformatics Node, Barcelona, Spain;

Department of Biochemistry and Molecular Biology, University of Barcelona,

Barcelona, Spain.

Despite of the variety of available Web services registries specially aimed at

Life Sciences, their scope is usually restricted to a limited set of well-defined

types of services. While dedicated registries are generally tied to a particular

format, general-purpose ones are more adherent to standards and usually rely on

Web Service Definition Language (WSDL). Although WSDL is quite flexible to

support common Web services types, its lack of semantic expressiveness led to

various initiatives to describe Web services via ontology languages.

Nevertheless, WSDL 2.0 descriptions gained a standard representation based on Web

Ontology Language (OWL). BioSWR is a novel Web services registry that provides

standard Resource Description Framework (RDF) based Web services descriptions

along with the traditional WSDL based ones. The registry provides Web-based

interface for Web services registration, querying and annotation, and is also

accessible programmatically via Representational State Transfer (REST) API or

using a SPARQL Protocol and RDF Query Language. BioSWR server is located at

http://inb.bsc.es/BioSWR/and its code is available at

https://sourceforge.net/projects/bioswr/under the LGPL license.

DOI: 10.1371/journal.pone.0107889

PMCID: PMC4169436

PMID: 25233118 [Indexed for MEDLINE]

730. PLoS One. 2014 Sep 17;9(9):e107837. doi: 10.1371/journal.pone.0107837.

eCollection 2014.

Search for β2 adrenergic receptor ligands by virtual screening via grid computing

and investigation of binding modes by docking and molecular dynamics simulations.

Bai Q(1), Shao Y(2), Pan D(2), Zhang Y(3), Liu H(4), Yao X(5).

Author information:

(1)Department of Chemistry, Lanzhou University, Lanzhou, China; School of Basic

Medical Sciences, Lanzhou University, Lanzhou, China. (2)Department of Chemistry,

Lanzhou University, Lanzhou, China. (3)School of Information Science &

Engineering, Lanzhou University, Lanzhou, China. (4)School of Pharmacy, Lanzhou

University, Lanzhou, China. (5)Department of Chemistry, Lanzhou University,

Lanzhou, China; State Key Laboratory of Quality Research in Chinese Medicine,

Macau Institute for Applied Research in Medicine and Health, Macau University of

Science and Technology, Taipa, Macau, China.

We designed a program called MolGridCal that can be used to screen small molecule

database in grid computing on basis of JPPF grid environment. Based on MolGridCal

program, we proposed an integrated strategy for virtual screening and binding

mode investigation by combining molecular docking, molecular dynamics (MD)

simulations and free energy calculations. To test the effectiveness of

MolGridCal, we screened potential ligands for β2 adrenergic receptor (β2AR) from

a database containing 50,000 small molecules. MolGridCal can not only send tasks

to the grid server automatically, but also can distribute tasks using the

screensaver function. As for the results of virtual screening, the known agonist

BI-167107 of β2AR is ranked among the top 2% of the screened candidates,

indicating MolGridCal program can give reasonable results. To further study the

binding mode and refine the results of MolGridCal, more accurate docking and

scoring methods are used to estimate the binding affinity for the top three

molecules (agonist BI-167107, neutral antagonist alprenolol and inverse agonist

ICI 118,551). The results indicate agonist BI-167107 has the best binding

affinity. MD simulation and free energy calculation are employed to investigate

the dynamic interaction mechanism between the ligands and β2AR. The results show

that the agonist BI-167107 also has the lowest binding free energy. This study

can provide a new way to perform virtual screening effectively through

integrating molecular docking based on grid computing, MD simulations and free

energy calculations. The source codes of MolGridCal are freely available at

http://molgridcal.codeplex.com.

DOI: 10.1371/journal.pone.0107837

PMCID: PMC4168136

PMID: 25229694 [Indexed for MEDLINE]

731. PLoS One. 2014 Sep 17;9(9):e107676. doi: 10.1371/journal.pone.0107676.

eCollection 2014.

A new supervised over-sampling algorithm with application to protein-nucleotide

binding residue prediction.

Hu J(1), He X(1), Yu DJ(2), Yang XB(3), Yang JY(1), Shen HB(4).

Author information:

(1)School of Computer Science and Engineering, Nanjing University of Science and

Technology, Nanjing, Jiangsu, China. (2)School of Computer Science and

Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu,

China; Changshu Institute, Nanjing University of Science and Technology,

Changshu, Jiangsu, China. (3)School of Computer Science and Engineering, Nanjing

University of Science and Technology, Nanjing, Jiangsu, China; School of Computer

Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang,

Jiangsu, China. (4)Institute of Image Processing and Pattern Recognition,

Shanghai Jiao Tong University, Shanghai, China.

Protein-nucleotide interactions are ubiquitous in a wide variety of biological

processes. Accurately identifying interaction residues solely from protein

sequences is useful for both protein function annotation and drug design,

especially in the post-genomic era, as large volumes of protein data have not

been functionally annotated. Protein-nucleotide binding residue prediction is a

typical imbalanced learning problem, where binding residues are extremely fewer

in number than non-binding residues. Alleviating the severity of class imbalance

has been demonstrated to be a promising means of improving the prediction

performance of a machine-learning-based predictor for class imbalance problems.

However, little attention has been paid to the negative impact of class imbalance

on protein-nucleotide binding residue prediction. In this study, we propose a new

supervised over-sampling algorithm that synthesizes additional minority class

samples to address class imbalance. The experimental results from

protein-nucleotide interaction datasets demonstrate that the proposed supervised

over-sampling algorithm can relieve the severity of class imbalance and help to

improve prediction performance. Based on the proposed over-sampling algorithm, a

predictor, called TargetSOS, is implemented for protein-nucleotide binding

residue prediction. Cross-validation tests and independent validation tests

demonstrate the effectiveness of TargetSOS. The web-server and datasets used in

this study are freely available at

http://www.csbio.sjtu.edu.cn/bioinf/TargetSOS/.

DOI: 10.1371/journal.pone.0107676

PMCID: PMC4168127

PMID: 25229688 [Indexed for MEDLINE]

732. Bioinformatics. 2014 Sep 15;30(18):2668-9. doi: 10.1093/bioinformatics/btu350.

Epub 2014 May 28.

MoDPepInt: an interactive web server for prediction of modular domain-peptide

interactions.

Kundu K(1), Mann M(1), Costa F(1), Backofen R(2).

Author information:

(1)Bioinformatics Group, Department of Computer Science, 79110 Freiburg and

Centre for Biological Signalling Studies (BIOSS), 79104 Freiburg, University of

Freiburg, Germany. (2)Bioinformatics Group, Department of Computer Science, 79110

Freiburg and Centre for Biological Signalling Studies (BIOSS), 79104 Freiburg,

University of Freiburg, Germany Bioinformatics Group, Department of Computer

Science, 79110 Freiburg and Centre for Biological Signalling Studies (BIOSS),

79104 Freiburg, University of Freiburg, Germany.

MoDPepInt (Modular Domain Peptide Interaction) is a new easy-to-use web server

for the prediction of binding partners for modular protein domains. Currently, we

offer models for SH2, SH3 and PDZ domains via the tools SH2PepInt, SH3PepInt and

PDZPepInt, respectively. More specifically, our server offers predictions for 51

SH2 human domains and 69 SH3 human domains via single domain models, and

predictions for 226 PDZ domains across several species, via 43 multidomain

models. All models are based on support vector machines with different kernel

functions ranging from polynomial, to Gaussian, to advanced graph kernels. In

this way, we model non-linear interactions between amino acid residues. Results

were validated on manually curated datasets achieving competitive performance

against various state-of-the-art approaches.AVAILABILITY AND IMPLEMENTATION: The

MoDPepInt server is available under the URL

http://modpepint.informatik.uni-freiburg.de/.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu350

PMCID: PMC4155253

PMID: 24872426 [Indexed for MEDLINE]

733. Front Microbiol. 2014 Sep 15;5:482. doi: 10.3389/fmicb.2014.00482. eCollection

2014.

Ori-Finder 2, an integrated tool to predict replication origins in the archaeal

genomes.

Luo H(1), Zhang CT(1), Gao F(2).

Author information:

(1)Department of Physics, Tianjin University Tianjin, China. (2)Department of

Physics, Tianjin University Tianjin, China ; Key Laboratory of Systems

Bioengineering (Ministry of Education), Tianjin University Tianjin, China ;

SynBio Research Platform, Collaborative Innovation Center of Chemical Science and

Engineering Tianjin, China.

DNA replication is one of the most basic processes in all three domains of

cellular life. With the advent of the post-genomic era, the increasing number of

complete archaeal genomes has created an opportunity for exploration of the

molecular mechanisms for initiating cellular DNA replication by in vivo

experiments as well as in silico analysis. However, the location of replication

origins (oriCs) in many sequenced archaeal genomes remains unknown. We present a

web-based tool Ori-Finder 2 to predict oriCs in the archaeal genomes

automatically, based on the integrated method comprising the analysis of base

composition asymmetry using the Z-curve method, the distribution of origin

recognition boxes identified by FIMO tool, and the occurrence of genes frequently

close to oriCs. The web server is also able to analyze the unannotated genome

sequences by integrating with gene prediction pipelines and BLAST software for

gene identification and function annotation. The result of the predicted oriCs is

displayed as an HTML table, which offers an intuitive way to browse the result in

graphical and tabular form. The software presented here is accurate for the

genomes with single oriC, but it does not necessarily find all the origins of

replication for the genomes with multiple oriCs. Ori-Finder 2 aims to become a

useful platform for the identification and analysis of oriCs in the archaeal

genomes, which would provide insight into the replication mechanisms in archaea.

The web server is freely available at http://tubic.tju.edu.cn/Ori-Finder2/.

DOI: 10.3389/fmicb.2014.00482

PMCID: PMC4164010

PMID: 25309521

734. PLoS One. 2014 Sep 15;9(9):e106542. doi: 10.1371/journal.pone.0106542.

eCollection 2014.

Random forest-based protein model quality assessment (RFMQA) using structural

features and potential energy terms.

Manavalan B(1), Lee J(1), Lee J(1).

Author information:

(1)Center for In Silico Protein Science, School of Computational Sciences, Korea

Institute for Advanced Study, Seoul, Korea.

Recently, predicting proteins three-dimensional (3D) structure from its sequence

information has made a significant progress due to the advances in computational

techniques and the growth of experimental structures. However, selecting good

models from a structural model pool is an important and challenging task in

protein structure prediction. In this study, we present the first application of

random forest based model quality assessment (RFMQA) to rank protein models using

its structural features and knowledge-based potential energy terms. The method

predicts a relative score of a model by using its secondary structure, solvent

accessibility and knowledge-based potential energy terms. We trained and tested

the RFMQA method on CASP8 and CASP9 targets using 5-fold cross-validation. The

correlation coefficient between the TM-score of the model selected by RFMQA

(TMRF) and the best server model (TMbest) is 0.945. We benchmarked our method on

recent CASP10 targets by using CASP8 and 9 server models as a training set. The

correlation coefficient and average difference between TMRF and TMbest over 95

CASP10 targets are 0.984 and 0.0385, respectively. The test results show that our

method works better in selecting top models when compared with other top

performing methods. RFMQA is available for download from

http://lee.kias.re.kr/RFMQA/RFMQA\_eval.tar.gz.

DOI: 10.1371/journal.pone.0106542

PMCID: PMC4164442

PMID: 25222008 [Indexed for MEDLINE]

735. PLoS One. 2014 Sep 12;9(9):e107504. doi: 10.1371/journal.pone.0107504.

eCollection 2014.

Vfold: a web server for RNA structure and folding thermodynamics prediction.

Xu X(1), Zhao P(1), Chen SJ(1).

Author information:

(1)Department of Physics and Department of Biochemistry, University of Missouri,

Columbia, Missouri, United States of America.

BACKGROUND: The ever increasing discovery of non-coding RNAs leads to

unprecedented demand for the accurate modeling of RNA folding, including the

predictions of two-dimensional (base pair) and three-dimensional all-atom

structures and folding stabilities. Accurate modeling of RNA structure and

stability has far-reaching impact on our understanding of RNA functions in human

health and our ability to design RNA-based therapeutic strategies.

RESULTS: The Vfold server offers a web interface to predict (a) RNA

two-dimensional structure from the nucleotide sequence, (b) three-dimensional

structure from the two-dimensional structure and the sequence, and (c) folding

thermodynamics (heat capacity melting curve) from the sequence. To predict the

two-dimensional structure (base pairs), the server generates an ensemble of

structures, including loop structures with the different intra-loop mismatches,

and evaluates the free energies using the experimental parameters for the base

stacks and the loop entropy parameters given by a coarse-grained RNA folding

model (the Vfold model) for the loops. To predict the three-dimensional

structure, the server assembles the motif scaffolds using structure templates

extracted from the known PDB structures and refines the structure using all-atom

energy minimization.

CONCLUSIONS: The Vfold-based web server provides a user friendly tool for the

prediction of RNA structure and stability. The web server and the source codes

are freely accessible for public use at "http://rna.physics.missouri.edu".

DOI: 10.1371/journal.pone.0107504

PMCID: PMC4162592

PMID: 25215508 [Indexed for MEDLINE]

736. Version 2. F1000Res. 2014 Sep 9 [revised 2014 Dec 1];3:214. doi:

10.12688/f1000research.5165.2. eCollection 2014.

ABS-Scan: In silico alanine scanning mutagenesis for binding site residues in

protein-ligand complex.

Anand P(1), Nagarajan D(1), Mukherjee S(2), Chandra N(1).

Author information:

(1)Department of Biochemistry, Indian Institute of Science, Bangalore, 560012,

India. (2)IISc Mathematics Initiative, Indian Institute of Science, Bangalore,

560012, India.

Most physiological processes in living systems are fundamentally regulated by

protein-ligand interactions. Understanding the process of ligand recognition by

proteins is a vital activity in molecular biology and biochemistry. It is well

known that the residues present at the binding site of the protein form pockets

that provide a conducive environment for recognition of specific ligands. In many

cases, the boundaries of these sites are not well defined. Here, we provide a

web-server to systematically evaluate important residues in the binding site of

the protein that contribute towards the ligand recognition through in silico

alanine-scanning mutagenesis experiments. Each of the residues present at the

binding site is computationally mutated to alanine. The ligand interaction energy

is computed for each mutant and the corresponding ΔΔG values are calculated by

comparing it to the wild type protein, thus evaluating individual residue

contributions towards ligand interaction. The server will thus provide a ranked

list of residues to the user in order to obtain loss-of-function mutations. This

web-tool can be freely accessed through the following address:

http://proline.biochem.iisc.ernet.in/abscan/.

DOI: 10.12688/f1000research.5165.2

PMCID: PMC4319546

PMID: 25685322

737. PLoS One. 2014 Sep 9;9(9):e104382. doi: 10.1371/journal.pone.0104382. eCollection

2014.

Accelerating translational research by clinically driven development of an

informatics platform--a case study.

Abugessaisa I(1), Saevarsdottir S(2), Tsipras G(1), Lindblad S(3), Sandin C(2),

Nikamo P(4), Ståhle M(4), Malmström V(2), Klareskog L(2), Tegnér J(1).

Author information:

(1)Unit of Computational Medicine, Department of Medicine, Center for Molecular

Medicine, Karolinska Institutet, Stockholm, Sweden. (2)Rheumatology Unit,

Department of Medicine, Center for Molecular Medicine, Karolinska Institutet,

Stockholm, Sweden. (3)Medical Management Center, Department of Learning,

Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden.

(4)Dermatology and Venereology Unit, Department of Medicine, Karolinska

Institutet, Karolinska University Hospital, Stockholm, Sweden.

Translational medicine is becoming increasingly dependent upon data generated

from health care, clinical research, and molecular investigations. This

increasing rate of production and diversity in data has brought about several

challenges, including the need to integrate fragmented databases, enable

secondary use of patient clinical data from health care in clinical research, and

to create information systems that clinicians and biomedical researchers can

readily use. Our case study effectively integrates requirements from the clinical

and biomedical researcher perspectives in a translational medicine setting. Our

three principal achievements are (a) a design of a user-friendly web-based system

for management and integration of clinical and molecular databases, while

adhering to proper de-identification and security measures; (b) providing a

real-world test of the system functionalities using clinical cohorts; and (c)

system integration with a clinical decision support system to demonstrate system

interoperability. We engaged two active clinical cohorts, 747 psoriasis patients

and 2001 rheumatoid arthritis patients, to demonstrate efficient query

possibilities across the data sources, enable cohort stratification, extract

variation in antibody patterns, study biomarker predictors of treatment response

in RA patients, and to explore metabolic profiles of psoriasis patients. Finally,

we demonstrated system interoperability by enabling integration with an

established clinical decision support system in health care. To assure the

usefulness and usability of the system, we followed two approaches. First, we

created a graphical user interface supporting all user interactions. Secondly we

carried out a system performance evaluation study where we measured the average

response time in seconds for active users, http errors, and kilobits per second

received and sent. The maximum response time was found to be 0.12 seconds; no

server or client errors of any kind were detected. In conclusion, the system can

readily be used by clinicians and biomedical researchers in a translational

medicine setting.

DOI: 10.1371/journal.pone.0104382

PMCID: PMC4159182

PMID: 25203647 [Indexed for MEDLINE]

738. BMC Bioinformatics. 2014 Sep 8;15:298. doi: 10.1186/1471-2105-15-298.

nDNA-Prot: identification of DNA-binding proteins based on unbalanced

classification.

Song L, Li D, Zeng X, Wu Y, Guo L(1), Zou Q.

Author information:

(1)School of Information Science and Technology, Xiamen University, Xiamen,

Fujian 361005, China. gl8008@163.com.

BACKGROUND: DNA-binding proteins are vital for the study of cellular processes.

In recent genome engineering studies, the identification of proteins with certain

functions has become increasingly important and needs to be performed rapidly and

efficiently. In previous years, several approaches have been developed to improve

the identification of DNA-binding proteins. However, the currently available

resources are insufficient to accurately identify these proteins. Because of

this, the previous research has been limited by the relatively unbalanced

accuracy rate and the low identification success of the current methods.

RESULTS: In this paper, we explored the practicality of modelling DNA binding

identification and simultaneously employed an ensemble classifier, and a new

predictor (nDNA-Prot) was designed. The presented framework is comprised of two

stages: a 188-dimension feature extraction method to obtain the protein structure

and an ensemble classifier designated as imDC. Experiments using different

datasets showed that our method is more successful than the traditional methods

in identifying DNA-binding proteins. The identification was conducted using a

feature that selected the minimum Redundancy and Maximum Relevance (mRMR). An

accuracy rate of 95.80% and an Area Under the Curve (AUC) value of 0.986 were

obtained in a cross validation. A test dataset was tested in our method and

resulted in an 86% accuracy, versus a 76% using iDNA-Prot and a 68% accuracy

using DNA-Prot.

CONCLUSIONS: Our method can help to accurately identify DNA-binding proteins, and

the web server is accessible at http://datamining.xmu.edu.cn/~songli/nDNA. In

addition, we also predicted possible DNA-binding protein sequences in all of the

sequences from the UniProtKB/Swiss-Prot database.

DOI: 10.1186/1471-2105-15-298

PMCID: PMC4165999

PMID: 25196432 [Indexed for MEDLINE]

739. BMC Bioinformatics. 2014 Sep 5;15:297. doi: 10.1186/1471-2105-15-297.

Enhancing protein-vitamin binding residues prediction by multiple heterogeneous

subspace SVMs ensemble.

Yu DJ(1), Hu J, Yan H, Yang XB, Yang JY, Shen HB.

Author information:

(1)School of Computer Science and Engineering, Nanjing University of Science and

Technology, Xiaolingwei 200, Nanjing 210094, China. njyudj@njust.edu.cn.

BACKGROUND: Vitamins are typical ligands that play critical roles in various

metabolic processes. The accurate identification of the vitamin-binding residues

solely based on a protein sequence is of significant importance for the

functional annotation of proteins, especially in the post-genomic era, when large

volumes of protein sequences are accumulating quickly without being functionally

annotated.

RESULTS: In this paper, a new predictor called TargetVita is designed and

implemented for predicting protein-vitamin binding residues using protein

sequences. In TargetVita, features derived from the position-specific scoring

matrix (PSSM), predicted protein secondary structure, and vitamin binding

propensity are combined to form the original feature space; then, several feature

subspaces are selected by performing different feature selection methods.

Finally, based on the selected feature subspaces, heterogeneous SVMs are trained

and then ensembled for performing prediction.

CONCLUSIONS: The experimental results obtained with four separate vitamin-binding

benchmark datasets demonstrate that the proposed TargetVita is superior to the

state-of-the-art vitamin-specific predictor, and an average improvement of 10% in

terms of the Matthews correlation coefficient (MCC) was achieved over independent

validation tests. The TargetVita web server and the datasets used are freely

available for academic use at http://csbio.njust.edu.cn/bioinf/TargetVita or

http://www.csbio.sjtu.edu.cn/bioinf/TargetVita.

DOI: 10.1186/1471-2105-15-297

PMCID: PMC4261549

PMID: 25189131 [Indexed for MEDLINE]

740. J Proteome Res. 2014 Sep 5;13(9):4184-91. doi: 10.1021/pr500525e. Epub 2014 Jul

30.

Online quantitative proteomics p-value calculator for permutation-based

statistical testing of peptide ratios.

Chen D(1), Shah A, Nguyen H, Loo D, Inder KL, Hill MM.

Author information:

(1)School of Information and Communication Technology, Griffith University , 170

Kessels Road, Nathan, Brisbane, Queensland 4111, Australia.

The utility of high-throughput quantitative proteomics to identify differentially

abundant proteins en-masse relies on suitable and accessible statistical

methodology, which remains mostly an unmet need. We present a free web-based

tool, called Quantitative Proteomics p-value Calculator (QPPC), designed for

accessibility and usability by proteomics scientists and biologists. Being an

online tool, there is no requirement for software installation. Furthermore, QPPC

accepts generic peptide ratio data generated by any mass spectrometer and

database search engine. Importantly, QPPC utilizes the permutation test that we

recently found to be superior to other methods for analysis of peptide ratios

because it does not assume normal distributions.1 QPPC assists the user in

selecting significantly altered proteins based on numerical fold change, or

standard deviation from the mean or median, together with the permutation

p-value. Output is in the form of comma separated values files, along with

graphical visualization using volcano plots and histograms. We evaluate the

optimal parameters for use of QPPC, including the permutation level and the

effect of outlier and contaminant peptides on p-value variability. The optimal

parameters defined are deployed as default for the web-tool at

http://qppc.di.uq.edu.au/ .

DOI: 10.1021/pr500525e

PMID: 25058807 [Indexed for MEDLINE]

741. PLoS One. 2014 Sep 3;9(9):e106691. doi: 10.1371/journal.pone.0106691. eCollection

2014.

iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid

distance-pairs and reduced alphabet profile into the general pseudo amino acid

composition.

Liu B(1), Xu J(2), Lan X(3), Xu R(4), Zhou J(2), Wang X(4), Chou KC(5).

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology

Shenzhen Graduate School, Shenzhen, Guangdong, China; Key Laboratory of Network

Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, Guangdong, China; Shanghai Key Laboratory of

Intelligent Information Processing, Shanghai, China; Gordon Life Science

Institute, Belmont, Massachusetts, United States of America. (2)School of

Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate

School, Shenzhen, Guangdong, China. (3)Stanford University, Stanford, California,

United States of America. (4)School of Computer Science and Technology, Harbin

Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China; Key

Laboratory of Network Oriented Intelligent Computation, Harbin Institute of

Technology Shenzhen Graduate School, Shenzhen, Guangdong, China. (5)Gordon Life

Science Institute, Belmont, Massachusetts, United States of America; Center of

Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University,

Jeddah, Saudi Arabia.

Playing crucial roles in various cellular processes, such as recognition of

specific nucleotide sequences, regulation of transcription, and regulation of

gene expression, DNA-binding proteins are essential ingredients for both

eukaryotic and prokaryotic proteomes. With the avalanche of protein sequences

generated in the postgenomic age, it is a critical challenge to develop automated

methods for accurate and rapidly identifying DNA-binding proteins based on their

sequence information alone. Here, a novel predictor, called "iDNA-Prot|dis", was

established by incorporating the amino acid distance-pair coupling information

and the amino acid reduced alphabet profile into the general pseudo amino acid

composition (PseAAC) vector. The former can capture the characteristics of

DNA-binding proteins so as to enhance its prediction quality, while the latter

can reduce the dimension of PseAAC vector so as to speed up its prediction

process. It was observed by the rigorous jackknife and independent dataset tests

that the new predictor outperformed the existing predictors for the same purpose.

As a user-friendly web-server, iDNA-Prot|dis is accessible to the public at

http://bioinformatics.hitsz.edu.cn/iDNA-Prot\_dis/. Moreover, for the convenience

of the vast majority of experimental scientists, a step-by-step protocol guide is

provided on how to use the web-server to get their desired results without the

need to follow the complicated mathematic equations that are presented in this

paper just for the integrity of its developing process. It is anticipated that

the iDNA-Prot|dis predictor may become a useful high throughput tool for

large-scale analysis of DNA-binding proteins, or at the very least, play a

complementary role to the existing predictors in this regard.

DOI: 10.1371/journal.pone.0106691

PMCID: PMC4153653

PMID: 25184541 [Indexed for MEDLINE]

742. Bioinformatics. 2014 Sep 1;30(17):i489-96. doi: 10.1093/bioinformatics/btu459.

CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding

strand at CRISPR loci.

Alkhnbashi OS(1), Costa F(1), Shah SA(1), Garrett RA(1), Saunders SJ(1), Backofen

R(2).

Author information:

(1)Department of Computer Science, University of Freiburg, Georges-Köhler-Allee

106, 79110 Freiburg, Germany, Department of Biology, University of Copenhagen,

Archaea Centre, Ole Maaloes Vej 5, DK2200 Copenhagen, Denmark and BIOSS Centre

for Biological Signalling Studies, Cluster of Excellence, University of Freiburg,

Germany. (2)Department of Computer Science, University of Freiburg,

Georges-Köhler-Allee 106, 79110 Freiburg, Germany, Department of Biology,

University of Copenhagen, Archaea Centre, Ole Maaloes Vej 5, DK2200 Copenhagen,

Denmark and BIOSS Centre for Biological Signalling Studies, Cluster of

Excellence, University of Freiburg, Germany Department of Computer Science,

University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany,

Department of Biology, University of Copenhagen, Archaea Centre, Ole Maaloes Vej

5, DK2200 Copenhagen, Denmark and BIOSS Centre for Biological Signalling Studies,

Cluster of Excellence, University of Freiburg, Germany.

MOTIVATION: The discovery of CRISPR-Cas systems almost 20 years ago rapidly

changed our perception of the bacterial and archaeal immune systems. CRISPR loci

consist of several repetitive DNA sequences called repeats, inter-spaced by

stretches of variable length sequences called spacers. This CRISPR array is

transcribed and processed into multiple mature RNA species (crRNAs). A single

crRNA is integrated into an interference complex, together with CRISPR-associated

(Cas) proteins, to bind and degrade invading nucleic acids. Although existing

bioinformatics tools can recognize CRISPR loci by their characteristic

repeat-spacer architecture, they generally output CRISPR arrays of ambiguous

orientation and thus do not determine the strand from which crRNAs are processed.

Knowledge of the correct orientation is crucial for many tasks, including the

classification of CRISPR conservation, the detection of leader regions, the

identification of target sites (protospacers) on invading genetic elements and

the characterization of protospacer-adjacent motifs.

RESULTS: We present a fast and accurate tool to determine the crRNA-encoding

strand at CRISPR loci by predicting the correct orientation of repeats based on

an advanced machine learning approach. Both the repeat sequence and mutation

information were encoded and processed by an efficient graph kernel to learn

higher-order correlations. The model was trained and tested on curated data

comprising >4500 CRISPRs and yielded a remarkable performance of 0.95 AUC ROC

(area under the curve of the receiver operator characteristic). In addition, we

show that accurate orientation information greatly improved detection of

conserved repeat sequence families and structure motifs. We integrated

CRISPRstrand predictions into our CRISPRmap web server of CRISPR conservation and

updated the latter to version 2.0.

AVAILABILITY: CRISPRmap and CRISPRstrand are available at

http://rna.informatik.uni-freiburg.de/CRISPRmap.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu459

PMCID: PMC4147912

PMID: 25161238 [Indexed for MEDLINE]

743. Bioinformatics. 2014 Sep 1;30(17):i364-70. doi: 10.1093/bioinformatics/btu441.

Towards a piRNA prediction using multiple kernel fusion and support vector

machine.

Brayet J(1), Zehraoui F(2), Jeanson-Leh L(2), Israeli D(2), Tahi F(2).

Author information:

(1)IBISC EA 4526, UEVE/Genopole, IBGBI, 23 bv. de France, 91000 Evry, France and

Genethon, 1, bis rue de l'Internationale, 91002 Evry Cedex, France IBISC EA 4526,

UEVE/Genopole, IBGBI, 23 bv. de France, 91000 Evry, France and Genethon, 1, bis

rue de l'Internationale, 91002 Evry Cedex, France. (2)IBISC EA 4526,

UEVE/Genopole, IBGBI, 23 bv. de France, 91000 Evry, France and Genethon, 1, bis

rue de l'Internationale, 91002 Evry Cedex, France.

MOTIVATION: Piwi-interacting RNA (piRNA) is the most recently discovered and the

least investigated class of Argonaute/Piwi protein-interacting small non-coding

RNAs. The piRNAs are mostly known to be involved in protecting the genome from

invasive transposable elements. But recent discoveries suggest their involvement

in the pathophysiology of diseases, such as cancer. Their identification is

therefore an important task, and computational methods are needed. However, the

lack of conserved piRNA sequences and structural elements makes this

identification challenging and difficult.

RESULTS: In the present study, we propose a new modular and extensible machine

learning method based on multiple kernels and a support vector machine (SVM)

classifier for piRNA identification. Very few piRNA features are known to date.

The use of a multiple kernels approach allows editing, adding or removing piRNA

features that can be heterogeneous in a modular manner according to their

relevance in a given species. Our algorithm is based on a combination of the

previously identified features [sequence features (k-mer motifs and a uridine at

the first position) and piRNAs cluster feature] and a new telomere/centromere

vicinity feature. These features are heterogeneous, and the kernels allow to

unify their representation. The proposed algorithm, named piRPred, gives

promising results on Drosophila and Human data and outscores previously published

piRNA identification algorithms.

AVAILABILITY AND IMPLEMENTATION: piRPred is freely available to non-commercial

users on our Web server EvryRNA http://EvryRNA.ibisc.univ-evry.fr.

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu441

PMCID: PMC4147894

PMID: 25161221 [Indexed for MEDLINE]

744. Bioinformatics. 2014 Sep 1;30(17):2519-20. doi: 10.1093/bioinformatics/btu334.

Epub 2014 May 14.

RNASeqExpressionBrowser--a web interface to browse and visualize high-throughput

expression data.

Nussbaumer T(1), Kugler KG(1), Bader KC(1), Sharma S(1), Seidel M(1), Mayer

KF(1).

Author information:

(1)Institute for Bioinformatics and Systems Biology (IBIS), Helmholtz Center

Munich, National Research Center for Environment and Health, 85764 Neuherberg,

Germany.

MOTIVATION: RNA-seq techniques generate massive amounts of expression data.

Several pipelines (e.g. Tophat and Cufflinks) are broadly applied to analyse

these datasets. However, accessing and handling the analytical output remain

challenging for non-experts.

RESULTS: We present the RNASeqExpressionBrowser, an open-source web interface

that can be used to access the output from RNA-seq expression analysis packages

in different ways, as it allows browsing for genes by identifiers, annotations or

sequence similarity. Gene expression information can be loaded as long as it is

represented in a matrix-like format. Additionally, data can be made available by

setting up the tool on a public server. For demonstration purposes, we have set

up a version providing expression information from the barley genome.

AVAILABILITY AND IMPLEMENTATION: The source code and a show case are accessible

at http://mips.helmholtz-muenchen.de/plant/RNASeqExpressionBrowser/.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu334

PMID: 24833805 [Indexed for MEDLINE]

745. Bioinformatics. 2014 Sep 1;30(17):2508-10. doi: 10.1093/bioinformatics/btu335.

Epub 2014 May 14.

RiboMaker: computational design of conformation-based riboregulation.

Rodrigo G(1), Jaramillo A(2).

Author information:

(1)Institute of Systems and Synthetic Biology, CNRS - Université d'Évry Val

d'Essonne, F-91000 Évry, France and School of Life Sciences, University of

Warwick, Coventry CV4 7AL, United Kingdom. (2)Institute of Systems and Synthetic

Biology, CNRS - Université d'Évry Val d'Essonne, F-91000 Évry, France and School

of Life Sciences, University of Warwick, Coventry CV4 7AL, United Kingdom

Institute of Systems and Synthetic Biology, CNRS - Université d'Évry Val

d'Essonne, F-91000 Évry, France and School of Life Sciences, University of

Warwick, Coventry CV4 7AL, United Kingdom.

MOTIVATION: The ability to engineer control systems of gene expression is

instrumental for synthetic biology. Thus, bioinformatic methods that assist such

engineering are appealing because they can guide the sequence design and prevent

costly experimental screening. In particular, RNA is an ideal substrate to de

novo design regulators of protein expression by following sequence-to-function

models.

RESULTS: We have implemented a novel algorithm, RiboMaker, aimed at the

computational, automated design of bacterial riboregulation. RiboMaker reads the

sequence and structure specifications, which codify for a gene regulatory

behaviour, and optimizes the sequences of a small regulatory RNA and a

5'-untranslated region for an efficient intermolecular interaction. To this end,

it implements an evolutionary design strategy, where random mutations are

selected according to a physicochemical model based on free energies. The

resulting sequences can then be tested experimentally, providing a new tool for

synthetic biology, and also for investigating the riboregulation principles in

natural systems.

AVAILABILITY AND IMPLEMENTATION: Web server is available at

http://ribomaker.jaramillolab.org/. Source code, instructions and examples are

freely available for download at http://sourceforge.net/projects/ribomaker/.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu335

PMID: 24833802 [Indexed for MEDLINE]

746. Bioinformatics. 2014 Sep 1;30(17):2534-6. doi: 10.1093/bioinformatics/btu241.

Epub 2014 Apr 23.

GenCLiP 2.0: a web server for functional clustering of genes and construction of

molecular networks based on free terms.

Wang JH(1), Zhao LF(1), Lin P(1), Su XR(1), Chen SJ(1), Huang LQ(1), Wang HF(1),

Zhang H(1), Hu ZF(1), Yao KT(1), Huang ZX(1).

Author information:

(1)Cancer Institute, Key Laboratory of Zebrafish Modeling and Drug Screening for

Human Diseases of Guangdong Higher Education Institutes, Department of Cell

Biology, Southern Medical University, Guangzhou 510515, Guangzhou Biotechnology

Center, Guangzhou, 510630, School of Basic Medical Sciences, Network Center and

Department of Plastic Surgery, Nanfang Hospital, Southern Medical University,

Guangzhou 510515, China.

Identifying biological functions and molecular networks in a gene list and how

the genes may relate to various topics is of considerable value to biomedical

researchers. Here, we present a web-based text-mining server, GenCLiP 2.0, which

can analyze human genes with enriched keywords and molecular interactions.

Compared with other similar tools, GenCLiP 2.0 offers two unique features: (i)

analysis of gene functions with free terms (i.e. any terms in the literature)

generated by literature mining or provided by the user and (ii) accurate

identification and integration of comprehensive molecular interactions from

Medline abstracts, to construct molecular networks and subnetworks related to the

free terms.AVAILABILITY AND IMPLEMENTATION: http://ci.smu.edu.cn.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu241

PMID: 24764463 [Indexed for MEDLINE]

747. J Mol Graph Model. 2014 Sep;53:59-71. doi: 10.1016/j.jmgm.2014.06.003. Epub 2014

Jul 14.

A collaborative visual analytics suite for protein folding research.

Harvey W(1), Park IH(2), Rübel O(3), Pascucci V(4), Bremer PT(5), Li C(6), Wang

Y(7).

Author information:

(1)Department of Computer Science and Engineering, The Ohio State University,

Columbus, OH, United States. Electronic address: harveywi@cse.ohio-state.edu.

(2)Chemical Physics Program, The Ohio State University, Columbus, OH, United

States. (3)Visualization Group, Lawrence Berkeley National Laboratory, Berkeley,

CA, United States. (4)Scientific Computing and Imaging Institute, University of

Utah, Salt Lake City, UT, United States. (5)Center for Applied Scientific

Computing, Lawrence Livermore National Laboratory, Livermore, CA, United States.

(6)Chemical Physics Program and College of Pharmacy, The Ohio State University,

Columbus, OH, United States. (7)Department of Computer Science and Engineering,

The Ohio State University, Columbus, OH, United States. Electronic address:

yusu@cse.ohio-state.edu.

Molecular dynamics (MD) simulation is a crucial tool for understanding principles

behind important biochemical processes such as protein folding and molecular

interaction. With the rapidly increasing power of modern computers, large-scale

MD simulation experiments can be performed regularly, generating huge amounts of

MD data. An important question is how to analyze and interpret such massive and

complex data. One of the (many) challenges involved in analyzing MD simulation

data computationally is the high-dimensionality of such data. Given a massive

collection of molecular conformations, researchers typically need to rely on

their expertise and prior domain knowledge in order to retrieve certain

conformations of interest. It is not easy to make and test hypotheses as the data

set as a whole is somewhat "invisible" due to its high dimensionality. In other

words, it is hard to directly access and examine individual conformations from a

sea of molecular structures, and to further explore the entire data set. There is

also no easy and convenient way to obtain a global view of the data or its

various modalities of biochemical information. To this end, we present an

interactive, collaborative visual analytics tool for exploring massive,

high-dimensional molecular dynamics simulation data sets. The most important

utility of our tool is to provide a platform where researchers can easily and

effectively navigate through the otherwise "invisible" simulation data sets,

exploring and examining molecular conformations both as a whole and at individual

levels. The visualization is based on the concept of a topological landscape,

which is a 2D terrain metaphor preserving certain topological and geometric

properties of the high dimensional protein energy landscape. In addition to

facilitating easy exploration of conformations, this 2D terrain metaphor also

provides a platform where researchers can visualize and analyze various

properties (such as contact density) overlayed on the top of the 2D terrain.

Finally, the software provides a collaborative environment where multiple

researchers can assemble observations and biochemical events into storyboards and

share them in real time over the Internet via a client-server architecture. The

software is written in Scala and runs on the cross-platform Java Virtual Machine.

Binaries and source code are available at http://www.aylasoftware.org and have

been released under the GNU General Public License.

Copyright © 2014 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.jmgm.2014.06.003

PMID: 25068440 [Indexed for MEDLINE]

748. Bioinformation. 2014 Aug 30;10(8):551-4. doi: 10.6026/97320630010551. eCollection

2014.

GSIT: An integrated web-tool for identification of genomic signatures in highly

similar DNA sequences.

Tuteja A(1), Joshi K(2), Subodh S(3), Kaur N(4).

Author information:

(1)Institute of Molecular Medicine, New Delhi ; Amity Institute of Biotechnology,

Amity University, Noida, India ; Authors contributed equally. (2)Institute of

Molecular Medicine, New Delhi ; Faculty of Technology and Engineering, The

Maharaja Sayajirao University of Baroda, Vadodara ; Authors contributed equally.

(3)Institute of Molecular Medicine, New Delhi ; Council of Scientific and

Industrial Research-Open Source Drug Discovery Unit, New Delhi. (4)Amity

Institute of Biotechnology, Amity University, Noida, India.

Accurate identification and characterization of infectious agent and its subtype

is essential for efficient treatment of infectious diseases on a target

population of patients. Comparative biology of microbial populations in vitro and

in vivo can identify signatures that may be used to develop and improve

diagnostic procedures. Here we report Genomic Signature Identification Tool

(GSIT) a web based tool for identification and validation of genomic signatures

in a group of similar DNA sequences of microorganisms. GSIT uses multiple

sequence alignment to identify the unique base sites and scores them for

inclusion as genomic signature for the particular strain. GSIT is a web based

tool where the front-end in designed using HTML/CSS and Javascript, while

back-end is run using CGI-Perl.AVAILABILITY: The server is freely available at

the http://genome-sign.net/gsit.

DOI: 10.6026/97320630010551

PMCID: PMC4166778

PMID: 25258494

749. J Chem Inf Model. 2014 Aug 25;54(8):2380-90. doi: 10.1021/ci5002197. Epub 2014

Aug 6.

Markov logic networks for optical chemical structure recognition.

Frasconi P(1), Gabbrielli F, Lippi M, Marinai S.

Author information:

(1)Dipartimento di Ingegneria dell'Informazione, Università degli Studi di

Firenze , Via di Santa Marta, 3, 50139 Firenze, Italy.

Optical chemical structure recognition is the problem of converting a bitmap

image containing a chemical structure formula into a standard structured

representation of the molecule. We introduce a novel approach to this problem

based on the pipelined integration of pattern recognition techniques with

probabilistic knowledge representation and reasoning. Basic entities and

relations (such as textual elements, points, lines, etc.) are first extracted by

a low-level processing module. A probabilistic reasoning engine based on Markov

logic, embodying chemical and graphical knowledge, is subsequently used to refine

these pieces of information. An annotated connection table of atoms and bonds is

finally assembled and converted into a standard chemical exchange format. We

report a successful evaluation on two large image data sets, showing that the

method compares favorably with the current state-of-the-art, especially on

degraded low-resolution images. The system is available as a web server at

http://mlocsr.dinfo.unifi.it.

DOI: 10.1021/ci5002197

PMID: 25068386 [Indexed for MEDLINE]

750. BMC Bioinformatics. 2014 Aug 20;15:282. doi: 10.1186/1471-2105-15-282.

Designing of peptides with desired half-life in intestine-like environment.

Sharma A, Singla D, Rashid M, Raghava GP(1).

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39A,

Chandigarh, India. raghava@imtech.res.in.

BACKGROUND: In past, a number of peptides have been reported to possess highly

diverse properties ranging from cell penetrating, tumor homing, anticancer,

anti-hypertensive, antiviral to antimicrobials. Owing to their excellent

specificity, low-toxicity, rich chemical diversity and availability from natural

sources, FDA has successfully approved a number of peptide-based drugs and

several are in various stages of drug development. Though peptides are proven

good drug candidates, their usage is still hindered mainly because of their high

susceptibility towards proteases degradation. We have developed an in silico

method to predict the half-life of peptides in intestine-like environment and to

design better peptides having optimized physicochemical properties and half-life.

RESULTS: In this study, we have used 10mer (HL10) and 16mer (HL16) peptides

dataset to develop prediction models for peptide half-life in intestine-like

environment. First, SVM based models were developed on HL10 dataset which

achieved maximum correlation R/R2 of 0.57/0.32, 0.68/0.46, and 0.69/0.47 using

amino acid, dipeptide and tripeptide composition, respectively. Secondly, models

developed on HL16 dataset showed maximum R/R2 of 0.91/0.82, 0.90/0.39, and

0.90/0.31 using amino acid, dipeptide and tripeptide composition, respectively.

Furthermore, models that were developed on selected features, achieved a

correlation (R) of 0.70 and 0.98 on HL10 and HL16 dataset, respectively.

Preliminary analysis suggests the role of charged residue and amino acid size in

peptide half-life/stability. Based on above models, we have developed a web

server named HLP (Half Life Prediction), for predicting and designing peptides

with desired half-life. The web server provides three facilities; i) half-life

prediction, ii) physicochemical properties calculation and iii) designing mutant

peptides.

CONCLUSION: In summary, this study describes a web server 'HLP' that has been

developed for assisting scientific community for predicting intestinal half-life

of peptides and to design mutant peptides with better half-life and

physicochemical properties. HLP models were trained using a dataset of peptides

whose half-lives have been determined experimentally in crude intestinal

proteases preparation. Thus, HLP server will help in designing peptides

possessing the potential to be administered via oral route

(http://www.imtech.res.in/raghava/hlp/).

DOI: 10.1186/1471-2105-15-282

PMCID: PMC4150950

PMID: 25141912 [Indexed for MEDLINE]

751. Database (Oxford). 2014 Aug 14;2014. pii: bau079. doi: 10.1093/database/bau079.

Print 2014.

PlantCAZyme: a database for plant carbohydrate-active enzymes.

Ekstrom A(1), Taujale R(1), McGinn N(1), Yin Y(2).

Author information:

(1)Department of Computer Science and Department of Biological Sciences, Northern

Illinois University, DeKalb, IL 60115, USA. (2)Department of Computer Science and

Department of Biological Sciences, Northern Illinois University, DeKalb, IL

60115, USA yyin@niu.edu.

PlantCAZyme is a database built upon dbCAN (database for automated carbohydrate

active enzyme annotation), aiming to provide pre-computed sequence and annotation

data of carbohydrate active enzymes (CAZymes) to plant carbohydrate and bioenergy

research communities. The current version contains data of 43,790 CAZymes of 159

protein families from 35 plants (including angiosperms, gymnosperms, lycophyte

and bryophyte mosses) and chlorophyte algae with fully sequenced genomes. Useful

features of the database include: (i) a BLAST server and a HMMER server that

allow users to search against our pre-computed sequence data for annotation

purpose, (ii) a download page to allow batch downloading data of a specific

CAZyme family or species and (iii) protein browse pages to provide an easy access

to the most comprehensive sequence and annotation data.DATABASE URL:

http://cys.bios.niu.edu/plantcazyme/

© The Author(s) 2014. Published by Oxford University Press.

DOI: 10.1093/database/bau079

PMCID: PMC4132414

PMID: 25125445 [Indexed for MEDLINE]

752. PLoS One. 2014 Aug 14;9(8):e105018. doi: 10.1371/journal.pone.0105018.

eCollection 2014.

iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo

amino acid composition.

Xu Y(1), Wen X(1), Wen LS(2), Wu LY(3), Deng NY(4), Chou KC(5).

Author information:

(1)Department of Information and Computer Science, University of Science and

Technology Beijing, Beijing, China. (2)College of Sciences, Liaoning Shiyou

University, FuShun, China. (3)Institute of Applied Mathematics, Academy of

Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China.

(4)College of Science, China Agricultural University, Beijing, China. (5)Center

of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University,

Jeddah, Saudi Arabia; Gordon Life Science Institute, Boston, Massachusetts,

United States of America.

Nitrotyrosine is one of the post-translational modifications (PTMs) in proteins

that occurs when their tyrosine residue is nitrated. Compared with healthy

people, a remarkably increased level of nitrotyrosine is detected in those

suffering from rheumatoid arthritis, septic shock, and coeliac disease. Given an

uncharacterized protein sequence that contains many tyrosine residues, which one

of them can be nitrated and which one cannot? This is a challenging problem, not

only directly related to in-depth understanding the PTM's mechanism but also to

the nitrotyrosine-based drug development. Particularly, with the avalanche of

protein sequences generated in the postgenomic age, it is highly desired to

develop a high throughput tool in this regard. Here, a new predictor called

"iNitro-Tyr" was developed by incorporating the position-specific dipeptide

propensity into the general pseudo amino acid composition for discriminating the

nitrotyrosine sites from non-nitrotyrosine sites in proteins. It was demonstrated

via the rigorous jackknife tests that the new predictor not only can yield higher

success rate but also is much more stable and less noisy. A web-server for

iNitro-Tyr is accessible to the public at http://app.aporc.org/iNitro-Tyr/. For

the convenience of most experimental scientists, we have further provided a

protocol of step-by-step guide, by which users can easily get their desired

results without the need to follow the complicated mathematics that were

presented in this paper just for the integrity of its development process. It has

not escaped our notice that the approach presented here can be also used to deal

with the other PTM sites in proteins.

DOI: 10.1371/journal.pone.0105018

PMCID: PMC4133382

PMID: 25121969 [Indexed for MEDLINE]

753. BMC Bioinformatics. 2014 Aug 13;15:277. doi: 10.1186/1471-2105-15-277.

CRF-based models of protein surfaces improve protein-protein interaction site

predictions.

Dong Z, Wang K, Dang TK, Gültas M, Welter M, Wierschin T, Stanke M, Waack S(1).

Author information:

(1)Institute of Computer Science, University of Göttingen, Goldschmidtstr, 7,

37077 Göttingen, Germany. waack@informatik.uni-goettingen.de.

BACKGROUND: The identification of protein-protein interaction sites is a

computationally challenging task and important for understanding the biology of

protein complexes. There is a rich literature in this field. A broad class of

approaches assign to each candidate residue a real-valued score that measures how

likely it is that the residue belongs to the interface. The prediction is

obtained by thresholding this score.Some probabilistic models classify the

residues on the basis of the posterior probabilities. In this paper, we introduce

pairwise conditional random fields (pCRFs) in which edges are not restricted to

the backbone as in the case of linear-chain CRFs utilized by Li et al. (2007). In

fact, any 3D-neighborhood relation can be modeled. On grounds of a generalized

Viterbi inference algorithm and a piecewise training process for pCRFs, we

demonstrate how to utilize pCRFs to enhance a given residue-wise score-based

protein-protein interface predictor on the surface of the protein under study.

The features of the pCRF are solely based on the interface predictions scores of

the predictor the performance of which shall be improved.

RESULTS: We performed three sets of experiments with synthetic scores assigned to

the surface residues of proteins taken from the data set PlaneDimers compiled by

Zellner et al. (2011), from the list published by Keskin et al. (2004) and from

the very recent data set due to Cukuroglu et al. (2014). That way we demonstrated

that our pCRF-based enhancer is effective given the interface residue score

distribution and the non-interface residue score are unimodal.Moreover, the

pCRF-based enhancer is also successfully applicable, if the distributions are

only unimodal over a certain sub-domain. The improvement is then restricted to

that domain. Thus we were able to improve the prediction of the PresCont server

devised by Zellner et al. (2011) on PlaneDimers.

CONCLUSIONS: Our results strongly suggest that pCRFs form a methodological

framework to improve residue-wise score-based protein-protein interface

predictors given the scores are appropriately distributed. A prototypical

implementation of our method is accessible at

http://ppicrf.informatik.uni-goettingen.de/index.html.

DOI: 10.1186/1471-2105-15-277

PMCID: PMC4150965

PMID: 25124108 [Indexed for MEDLINE]

754. J Theor Biol. 2014 Aug 7;354:48-53. doi: 10.1016/j.jtbi.2014.03.026. Epub 2014

Mar 24.

TIBS: a web database to browse gene expression in irritable bowel syndrome.

Yan J(1), Xu Y(2), Hu B(3), Alnajm S(4), Liu L(5), Lu Y(6), Sun Z(7), Cheng F(8).

Author information:

(1)Institute of First Clinical Medicine, Nanjing University of Chinese Medicine,

Nanjing 210046, Jiangsu, People Republic of China. Electronic address:

yanjing0513@126.com. (2)Department of Pharmaceutical Science, College of

Pharmacy, University of South Florida, Tampa 33612, USA. Electronic address:

yxu@mail.usf.edu. (3)Essenology Software Inc., Wuxi 214187, People Republic of

China. Electronic address: brianwhu@yahoo.com. (4)Department of Biology, Faculty

of Mathematics and Natural Sciences, University of Cologne, Cologne 50674,

Germany. Electronic address: salnajm@smail.uni-koeln.de. (5)Institute of First

Clinical Medicine, Nanjing University of Chinese Medicine, Nanjing 210046,

Jiangsu, People Republic of China. Electronic address: lina\_liu606@126.com.

(6)Department of Pharmacology, College of Pharmacy, Nanjing University of Chinese

Medicine, Nanjing 210046, People Republic of China. Electronic address:

profyinlu@163.com. (7)Institute of First Clinical Medicine, Nanjing University of

Chinese Medicine, Nanjing 210046, Jiangsu, People Republic of China. Electronic

address: pr\_zhiguangsun@163.com. (8)Department of Pharmaceutical Science, College

of Pharmacy, University of South Florida, Tampa 33612, USA; Department of

Department of Epidemiology and Biostatistics, College of Public Health,

University of South Florida, Tampa 33612, USA. Electronic address:

fcheng1@health.usf.edu.

Irritable bowel syndrome (IBS) is a chronic functional gastrointestinal disorder.

Its symptoms include chronic abdominal pain, bloating gas, diarrhea and

constipation. Many IBS patients also have psychological symptoms like depression

or anxiety. These unpleasant symptoms significantly lower patients׳ quality of

life. The prevalence of IBS in Europe and North America is about 10-15% of the

population, which makes IBS a disorder with a high social cost. The

pathophysiology of IBS is considered to be multifactorial and the exact cause of

the disease remains poorly understood. Recently, a genome-wide expression

microarray technique has been applied to investigate the possible mechanisms of

IBS. However, a user-friendly database that allows scientists without

bioinformatics background to query gene expression levels in these data sets and

compare gene expression patterns across different tissues has not yet been

established. Therefore, we have integrated four public expression microarray data

(320 samples) from the Gene Expression Omnibus (GEO) and ArrayExpress databases

into an online database called Transcriptome of Irritable Bowel Syndrome (TIBS).

The gene expression change in IBS patients compared to healthy volunteers or UC

patients in jejunum, sigmoid colon, rectum, and descending colon can be queried

by gene symbols. Users can compare gene expression levels of IBS patients across

these tissues. Sex difference of gene expression in IBS patients was also shown

in the database. The current version of TIBS database contains 42,400 annotated

gene probe sets represented on the Affymetrix Human Genome U133 plus 2.0

platform. TIBS will be an invaluable resource for a better understanding of the

pathogenesis of IBS at the molecular level and for drug development. The TIBS

database is available online at http://www.chengfeng.info/tibs\_database.html.

Copyright © 2014 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2014.03.026

PMID: 24675620 [Indexed for MEDLINE]

755. Biophys Chem. 2014 Aug;192:10-9. doi: 10.1016/j.bpc.2014.05.002. Epub 2014 May

29.

Prediction of fatty acid-binding residues on protein surfaces with

three-dimensional probability distributions of interacting atoms.

Mahalingam R(1), Peng HP(2), Yang AS(3).

Author information:

(1)Genomics Research Center, Academia Sinica, Taipei 115, Taiwan. Electronic

address: rajasekaran.mahalingam@case.edu. (2)Genomics Research Center, Academia

Sinica, Taipei 115, Taiwan; Institute of Biomedical Informatics, National

Yang-Ming University, Taipei 11221, Taiwan; Bioinformatics Program, Taiwan

International Graduate Program, Institute of Information Science, Academia

Sinica, Taipei 115, Taiwan. (3)Genomics Research Center, Academia Sinica, Taipei

115, Taiwan. Electronic address: yangas@gate.sinica.edu.tw.

Protein-fatty acid interaction is vital for many cellular processes and

understanding this interaction is important for functional annotation as well as

drug discovery. In this work, we present a method for predicting the fatty acid

(FA)-binding residues by using three-dimensional probability density

distributions of interacting atoms of FAs on protein surfaces which are derived

from the known protein-FA complex structures. A machine learning algorithm was

established to learn the characteristic patterns of the probability density maps

specific to the FA-binding sites. The predictor was trained with five-fold cross

validation on a non-redundant training set and then evaluated with an independent

test set as well as on holo-apo pair's dataset. The results showed good accuracy

in predicting the FA-binding residues. Further, the predictor developed in this

study is implemented as an online server which is freely accessible at the

following website, http://ismblab.genomics.sinica.edu.tw/.

Copyright © 2014 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.bpc.2014.05.002

PMID: 24934883 [Indexed for MEDLINE]

756. Mol Biosyst. 2014 Aug;10(8):2229-35. doi: 10.1039/c4mb00316k.

Identification of bacteriophage virion proteins by the ANOVA feature selection

and analysis.

Ding H(1), Feng PM, Chen W, Lin H.

Author information:

(1)Key Laboratory for Neuro-Information of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu 610054, China. hding@uestc.edu.cn

hlin@uestc.edu.cn.

The bacteriophage virion proteins play extremely important roles in the fate of

host bacterial cells. Accurate identification of bacteriophage virion proteins is

very important for understanding their functions and clarifying the lysis

mechanism of bacterial cells. In this study, a new sequence-based method was

developed to identify phage virion proteins. In the new method, the protein

sequences were initially formulated by the g-gap dipeptide compositions.

Subsequently, the analysis of variance (ANOVA) with incremental feature selection

(IFS) was used to search for the optimal feature set. It was observed that, in

jackknife cross-validation, the optimal feature set including 160 optimized

features can produce the maximum accuracy of 85.02%. By performing feature

analysis, we found that the correlation between two amino acids with one gap was

more important than other correlations for phage virion protein prediction and

that some of the 1-gap dipeptides were important and mainly contributed to the

virion protein prediction. This analysis will provide novel insights into the

function of phage virion proteins. On the basis of the proposed method, an online

web-server, PVPred, was established and can be freely accessed from the website

(http://lin.uestc.edu.cn/server/PVPred). We believe that the PVPred will become a

powerful tool to study phage virion proteins and to guide the related

experimental validations.

DOI: 10.1039/c4mb00316k

PMID: 24931825 [Indexed for MEDLINE]

757. Mol Biosyst. 2014 Aug;10(8):2031-42. doi: 10.1039/c4mb00289j.

IndividualizedPath: identifying genetic alterations contributing to the

dysfunctional pathways in glioblastoma individuals.

Ping Y(1), Zhang H, Deng Y, Wang L, Zhao H, Pang L, Fan H, Xu C, Li F, Zhang Y,

Gong Y, Xiao Y, Li X.

Author information:

(1)College of Bioinformatics Science and Technology, Harbin Medical University,

Harbin 150081, China. xiaoyun@ems.hrbmu.edu.cn lixia@hrbmu.edu.cn.

Due to the extensive complexity and high genetic heterogeneity of genetic

alterations in cancer, comprehensively depicting the molecular mechanisms of

cancer remains difficult. Characterizing personalized pathogenesis in cancer

individuals can help to reveal new details of the complex mechanisms. In this

study, we proposed an integrative method called IndividualizedPath to identify

genetic alterations and their downstream risk pathways from the perspective of

individuals through combining the DNA copy number, gene expression data and

topological structures of biological pathways. By applying the method to TCGA

glioblastoma multiforme (GBM) samples, we identified 394 gene-pathway pairs in

252 GBM individuals. We found that genes with copy number alterations showed high

heterogeneity across GBM individuals, whereas they affected relatively consistent

biological pathways. A global landscape of gene-pathway pairs showed that EGFR

linked with multiple cancer-related biological pathways confers the highest risk

of GBM. GBM individuals with MET-pathway pairs showed significantly shorter

survival times than those with only MET amplification. Importantly, we found that

the same risk pathways were affected by different genes in distinct groups of GBM

individuals with a significant pattern of mutual exclusivity. Similarly, GBM

subtype analysis revealed some subtype-specific gene-pathway pairs. In addition,

we found that some rare copy number alterations had a large effect on

contribution to numerous cancer-related pathways. In summary, our method offers

the possibility to identify personalized cancer mechanisms, which can be applied

to other types of cancer through the web server

(http://bioinfo.hrbmu.edu.cn/IndividualizedPath/).

DOI: 10.1039/c4mb00289j

PMID: 24911613 [Indexed for MEDLINE]

758. Pharmacoepidemiol Drug Saf. 2014 Aug;23(8):795-801. doi: 10.1002/pds.3561. Epub

2014 Feb 12.

AERS spider: an online interactive tool to mine statistical associations in

Adverse Event Reporting System.

Grigoriev I(1), zu Castell W, Tsvetkov P, Antonov AV.

Author information:

(1)Helmholtz Zentrum München GmbH, Department of Scientific Computing,

Neuherberg, Germany.

BACKGROUND: Exploration of the Adverse Event Reporting System (AERS) data by a

wide scientific community is limited due to several factors. First, AERS data

must be intensively preprocessed to be converted into analyzable format. Second,

application of the currently accepted disproportional reporting measures results

in false positive signals.

METHODS: We proposed a data mining strategy to improve hypothesis generation with

respect to potential associations.

RESULTS: By numerous examples, we illustrate that our strategy controls the false

positive signals. We implemented a free online tool, AERS spider

(www.chemoprofiling.org/AERS).

CONCLUSIONS: We believe that AERS spider would be a valuable tool for drug safety

experts.

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/pds.3561

PMID: 24677538 [Indexed for MEDLINE]

759. PLoS Comput Biol. 2014 Jul 31;10(7):e1003750. doi: 10.1371/journal.pcbi.1003750.

eCollection 2014.

A real-time all-atom structural search engine for proteins.

Gonzalez G(1), Hannigan B(1), DeGrado WF(1).

Author information:

(1)Cardiovascular Research Institute, University of California, San Francisco,

San Francisco, California, United States of America.

Protein designers use a wide variety of software tools for de novo design, yet

their repertoire still lacks a fast and interactive all-atom search engine. To

solve this, we have built the Suns program: a real-time, atomic search engine

integrated into the PyMOL molecular visualization system. Users build

atomic-level structural search queries within PyMOL and receive a stream of

search results aligned to their query within a few seconds. This instant feedback

cycle enables a new "designability"-inspired approach to protein design where the

designer searches for and interactively incorporates native-like fragments from

proven protein structures. We demonstrate the use of Suns to interactively build

protein motifs, tertiary interactions, and to identify scaffolds compatible with

hot-spot residues. The official web site and installer are located at

http://www.degradolab.org/suns/ and the source code is hosted at

https://github.com/godotgildor/Suns (PyMOL plugin, BSD license),

https://github.com/Gabriel439/suns-cmd (command line client, BSD license), and

https://github.com/Gabriel439/suns-search (search engine server, GPLv2 license).

DOI: 10.1371/journal.pcbi.1003750

PMCID: PMC4117414

PMID: 25079944 [Indexed for MEDLINE]

760. BMC Bioinformatics. 2014 Jul 28;15:254. doi: 10.1186/1471-2105-15-254.

ELM: enhanced lowest common ancestor based method for detecting a pathogenic

virus from a large sequence dataset.

Ueno K, Ishii A, Ito K(1).

Author information:

(1)Division of Bioinformatics, Research Center for Zoonosis Control, Hokkaido

University, Sapporo, Hokkaido 001-0020, Japan. itok@czc.hokudai.ac.jp.

BACKGROUND: Emerging viral diseases, most of which are caused by the transmission

of viruses from animals to humans, pose a threat to public health. Discovering

pathogenic viruses through surveillance is the key to preparedness for this

potential threat. Next generation sequencing (NGS) helps us to identify viruses

without the design of a specific PCR primer. The major task in NGS data analysis

is taxonomic identification for vast numbers of sequences. However, taxonomic

identification via a BLAST search against all the known sequences is a

computational bottleneck.

DESCRIPTION: Here we propose an enhanced lowest-common-ancestor based method

(ELM) to effectively identify viruses from massive sequence data. To reduce the

computational cost, ELM uses a customized database composed only of viral

sequences for the BLAST search. At the same time, ELM adopts a novel criterion to

suppress the rise in false positive assignments caused by the small database. As

a result, identification by ELM is more than 1,000 times faster than the

conventional methods without loss of accuracy.

CONCLUSIONS: We anticipate that ELM will contribute to direct diagnosis of viral

infections. The web server and the customized viral database are freely available

at http://bioinformatics.czc.hokudai.ac.jp/ELM/.

DOI: 10.1186/1471-2105-15-254

PMCID: PMC4124145

PMID: 25069839 [Indexed for MEDLINE]

761. J Chem Inf Model. 2014 Jul 28;54(7):2068-78. doi: 10.1021/ci500115j. Epub 2014

Jul 10.

Docking server for the identification of heparin binding sites on proteins.

Mottarella SE(1), Beglov D, Beglova N, Nugent MA, Kozakov D, Vajda S.

Author information:

(1)Program in Bioinformatics and ‡Department of Biomedical Engineering, Boston

University , 44 Cummington Street, Boston, Massachusetts 02215, United States.

Many proteins of widely differing functionality and structure are capable of

binding heparin and heparan sulfate. Since crystallizing protein-heparin

complexes for structure determination is generally difficult, computational

docking can be a useful approach for understanding specific interactions.

Previous studies used programs originally developed for docking small molecules

to well-defined pockets, rather than for docking polysaccharides to highly

charged shallow crevices that usually bind heparin. We have extended the program

PIPER and the automated protein-protein docking server ClusPro to heparin

docking. Using a molecular mechanics energy function for scoring and the fast

Fourier transform correlation approach, the method generates and evaluates close

to a billion poses of a heparin tetrasaccharide probe. The docked structures are

clustered using pairwise root-mean-square deviations as the distance measure. It

was shown that clustering of heparin molecules close to each other but having

different orientations and selecting the clusters with the highest protein-ligand

contacts reliably predicts the heparin binding site. In addition, the centers of

the five most populated clusters include structures close to the native

orientation of the heparin. These structures can provide starting points for

further refinement by methods that account for flexibility such as molecular

dynamics. The heparin docking method is available as an advanced option of the

ClusPro server at http://cluspro.bu.edu/ .

DOI: 10.1021/ci500115j

PMCID: PMC4184157

PMID: 24974889 [Indexed for MEDLINE]

762. JMIR Res Protoc. 2014 Jul 23;3(3):e31. doi: 10.2196/resprot.2599.

Intervention use and action planning in a web-based computer-tailored weight

management program for overweight adults: randomized controlled trial.

van Genugten L(1), van Empelen P, Oenema A.

Author information:

(1)Erasmus University Medical Center, Department of Public Health, Rotterdam,

Netherlands. l.vangenugten@erasmusmc.nl.

BACKGROUND: There are many online interventions aiming for health behavior change

but it is unclear how such interventions and specific planning tools are being

used.

OBJECTIVE: The aim of this study is to identify which user characteristics were

associated with use of an online, computer-tailored self-regulation intervention

aimed at prevention of weight gain; and to examine the quality of the goals and

action plans that were generated using the online planning tools.

METHODS: Data were obtained with a randomized controlled effect evaluation trial

in which the online computer-tailored intervention was compared to a website

containing generic information about prevention of weight gain. The tailored

intervention included self-regulation techniques such as personalized feedback,

goal setting, action planning, monitoring, and other techniques aimed at weight

management. Participants included 539 overweight adults (mean age 46.9 years,

mean body mass index [BMI] 28.03 kg/m(2), 31.2% male, 11% low education level)

recruited from the general population. Use of the intervention and its planning

tools were derived from server registration data. Physical activity, fat intake,

motivational factors, and self-regulation skills were self-reported at baseline.

Descriptive analyses and logistic regression analyses were used to analyze the

results.

RESULTS: Use of the tailored intervention decreased sharply after the first

modules. Visiting the first tailored intervention module was more likely among

participants with low levels of fat intake (OR 0.77, 95% CI 0.62-0.95) or

planning for change in PA (OR 0.23, 95% CI 0.05-0.97). Revisiting the

intervention was more likely among participants high in restrained eating (OR

2.45, 95% CI 1.12-5.43) or low in proactive coping skills for weight control (OR

0.28, 95% CI 0.10-0.76). The planning tools were used by 5%-55% of the

participants, but only 20%-75% of the plans were of good quality.

CONCLUSIONS: This study showed that psychological factors such as self-regulation

skills and action planning were associated with repeated use of an online,

computer-tailored self-regulation intervention aimed at prevention of weight gain

among adults being overweight. Use of the intervention was not optimal, with a

limited number of participants who visited all the intervention modules. The use

of the action and coping planning components of the intervention was mediocre and

the quality of the generated plans was low, especially for the coping plans. It

is important to identify how the use of action planning and coping planning

components in online interventions can be promoted and how the quality of plans

generated through these tools can be improved.

TRIAL REGISTRATION: Netherlands Trial Register: NTR1862;

http://www.trialregister.nl/trialreg/admin/rctview.asp?TC=1862 (Archived by

WebCite at http://www.webcitation.org/6QG1ZPIzZ).

DOI: 10.2196/resprot.2599

PMCID: PMC4129126

PMID: 25057122

763. BMC Bioinformatics. 2014 Jul 2;15:230. doi: 10.1186/1471-2105-15-230.

Osiris: accessible and reproducible phylogenetic and phylogenomic analyses within

the Galaxy workflow management system.

Oakley TH(1), Alexandrou MA, Ngo R, Pankey MS, Churchill CK, Chen W, Lopker KB.

Author information:

(1)Ecology, Evolution, and Marine Biology, University of California-Santa

Barbara, Santa Barbara, CA 93106, USA. oakley@lifesci.ucsb.edu.

BACKGROUND: Phylogenetic tools and 'tree-thinking' approaches increasingly

permeate all biological research. At the same time, phylogenetic data sets are

expanding at breakneck pace, facilitated by increasingly economical sequencing

technologies. Therefore, there is an urgent need for accessible, modular, and

sharable tools for phylogenetic analysis.

RESULTS: We developed a suite of wrappers for new and existing phylogenetics

tools for the Galaxy workflow management system that we call Osiris. Osiris and

Galaxy provide a sharable, standardized, modular user interface, and the ability

to easily create complex workflows using a graphical interface. Osiris enables

all aspects of phylogenetic analysis within Galaxy, including de novo assembly of

high throughput sequencing reads, ortholog identification, multiple sequence

alignment, concatenation, phylogenetic tree estimation, and post-tree comparative

analysis. The open source files are available on in the Bitbucket public

repository and many of the tools are demonstrated on a public web server

(http://galaxy-dev.cnsi.ucsb.edu/osiris/).

CONCLUSIONS: Osiris can serve as a foundation for other phylogenomic and

phylogenetic tool development within the Galaxy platform.

DOI: 10.1186/1471-2105-15-230

PMCID: PMC4227113

PMID: 24990571 [Indexed for MEDLINE]

764. Anal Biochem. 2014 Jul 1;456:53-60. doi: 10.1016/j.ab.2014.04.001. Epub 2014 Apr

13.

PseKNC: a flexible web server for generating pseudo K-tuple nucleotide

composition.

Chen W(1), Lei TY(2), Jin DC(2), Lin H(3), Chou KC(4).

Author information:

(1)School of Sciences, and Center for Genomics and Computational Biology, Hebei

United University, Tangshan 063000, China; Gordon Life Science Institute,

Belmont, MA 02478, USA. Electronic address: wchen@gordonlifescience.org.

(2)School of Sciences, and Center for Genomics and Computational Biology, Hebei

United University, Tangshan 063000, China. (3)Gordon Life Science Institute,

Belmont, MA 02478, USA; Key Laboratory for Neuro-Information of Ministry of

Education, Center of Bioinformatics, School of Life Science and Technology,

University of Electronic Science and Technology of China, Chengdu 610054, China.

Electronic address: hlin@uestc.edu.cn. (4)School of Sciences, and Center for

Genomics and Computational Biology, Hebei United University, Tangshan 063000,

China; Gordon Life Science Institute, Belmont, MA 02478, USA; Center of

Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University,

Jeddah 21589, Saudi Arabia. Electronic address: kcchou@gordonlifescience.org.

The pseudo oligonucleotide composition, or pseudo K-tuple nucleotide composition

(PseKNC), can be used to represent a DNA or RNA sequence with a discrete model or

vector yet still keep considerable sequence order information, particularly the

global or long-range sequence order information, via the physicochemical

properties of its constituent oligonucleotides. Therefore, the PseKNC approach

may hold very high potential for enhancing the power in dealing with many

problems in computational genomics and genome sequence analysis. However, dealing

with different DNA or RNA problems may need different kinds of PseKNC. Here, we

present a flexible and user-friendly web server for PseKNC (at

http://lin.uestc.edu.cn/pseknc/default.aspx) by which users can easily generate

many different modes of PseKNC according to their need by selecting various

parameters and physicochemical properties. Furthermore, for the convenience of

the vast majority of experimental scientists, a step-by-step guide is provided on

how to use the current web server to generate their desired PseKNC without the

need to follow the complicated mathematical equations, which are presented in

this article just for the integrity of PseKNC formulation and its development. It

is anticipated that the PseKNC web server will become a very useful tool in

computational genomics and genome sequence analysis.

Copyright © 2014 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2014.04.001

PMID: 24732113 [Indexed for MEDLINE]

765. Bioinformatics. 2014 Jul 1;30(13):1935-6. doi: 10.1093/bioinformatics/btu129.

Epub 2014 Mar 6.

Frag'r'Us: knowledge-based sampling of protein backbone conformations for de novo

structure-based protein design.

Bonet J(1), Segura J(1), Planas-Iglesias J(1), Oliva B(1), Fernandez-Fuentes

N(1).

Author information:

(1)Structural Bioinformatics Lab (GRIB), Universitat Pompeu Fabra, Barcelona,

Catalonia 08003, Spain, Department of Structure of Macromolecules, CNB-CSIC,

Cantoblanco, 28049 Madrid, Spain and Institute of Biological, Environmental and

Rural Science, Aberystwyth University, Aberystwyth, SY23 3DA Ceredigion, UK.

MOTIVATION: The remodeling of short fragment(s) of the protein backbone to

accommodate new function(s), fine-tune binding specificities or change/create

novel protein interactions is a common task in structure-based computational

design. Alternative backbone conformations can be generated de novo or by

redeploying existing fragments extracted from protein structures i.e.

knowledge-based. We present Frag'r'Us, a web server designed to sample

alternative protein backbone conformations in loop regions. The method relies on

a database of super secondary structural motifs called smotifs. Thus, sampling of

conformations reflects structurally feasible fragments compiled from existing

protein structures. Availability and implementation Frag'r'Us has been

implemented as web application and is available at

http://www.bioinsilico.org/FRAGRUS.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu129

PMID: 24603983 [Indexed for MEDLINE]

766. Brief Bioinform. 2014 Jul;15(4):471-83.

A review of genomic data warehousing systems.

Triplet T, Butler G.

To facilitate the integration and querying of genomics data, a number of generic

data warehousing frameworks have been developed. They differ in their design and

capabilities, as well as their intended audience. We provide a comprehensive and

quantitative review of those genomic data warehousing frameworks in the context

of large-scale systems biology. We reviewed in detail four genomic data

warehouses (BioMart, BioXRT, InterMine and PathwayTools) freely available to the

academic community. We quantified 20 aspects of the warehouses, covering the

accuracy of their responses, their computational requirements and development

efforts. Performance of the warehouses was evaluated under various hardware

configurations to help laboratories optimize hardware expenses. Each aspect of

the benchmark may be dynamically weighted by scientists using our online tool

BenchDW (http://warehousebenchmark.fungalgenomics.ca/benchmark/) to build custom

warehouse profiles and tailor our results to their specific needs.

DOI: 10.1093/bib/bbt031

PMID: 23673292 [Indexed for MEDLINE]

767. Brief Bioinform. 2014 Jul;15(4):648-59. doi: 10.1093/bib/bbs082. Epub 2013 Jan

31.

Pharmaco-miR: linking microRNAs and drug effects.

Rukov JL, Wilentzik R, Jaffe I, Vinther J, Shomron N.

MicroRNAs (miRNAs) are short regulatory RNAs that down-regulate gene expression.

They are essential for cell homeostasis and active in many disease states. A

major discovery is the ability of miRNAs to determine the efficacy of drugs,

which has given rise to the field of 'miRNA pharmacogenomics' through

'Pharmaco-miRs'. miRNAs play a significant role in pharmacogenomics by

down-regulating genes that are important for drug function. These interactions

can be described as triplet sets consisting of a miRNA, a target gene and a drug

associated with the gene. We have developed a web server which links miRNA

expression and drug function by combining data on miRNA targeting and

protein-drug interactions. miRNA targeting information derive from both

experimental data and computational predictions, and protein-drug interactions

are annotated by the Pharmacogenomics Knowledge base (PharmGKB). Pharmaco-miR's

input consists of miRNAs, genes and/or drug names and the output consists of

miRNA pharmacogenomic sets or a list of unique associated miRNAs, genes and

drugs. We have furthermore built a database, named Pharmaco-miR Verified Sets

(VerSe), which contains miRNA pharmacogenomic data manually curated from the

literature, can be searched and downloaded via Pharmaco-miR and informs on trends

and generalities published in the field. Overall, we present examples of how

Pharmaco-miR provides possible explanations for previously published

observations, including how the cisplatin and 5-fluorouracil resistance induced

by miR-148a may be caused by miR-148a targeting of the gene KIT. The information

is available at www.Pharmaco-miR.org.

© The Author 2013. Published by Oxford University Press.

DOI: 10.1093/bib/bbs082

PMCID: PMC4103536

PMID: 23376192 [Indexed for MEDLINE]

768. Clin Nutr Res. 2014 Jul;3(2):115-25. doi: 10.7762/cnr.2014.3.2.115. Epub 2014 Jul

29.

Development and Evaluation of a Web-based Computer-Assisted Personal Interview

System (CAPIS) for Open-ended Dietary Assessments among Koreans.

Shin S(1), Park E(2), Sun DH(3), You TK(3), Lee MJ(3), Hwang S(3), Paik HY(4),

Joung H(1).

Author information:

(1)Department of Public Health Nutrition Graduate School of Public Health, Seoul

National University, Seoul 151-742, Korea. ; Institute of Health and Environment,

Seoul National University, Seoul 151-742, Korea. (2)Department of Public Health

Nutrition Graduate School of Public Health, Seoul National University, Seoul

151-742, Korea. (3)Department of Computer Engineering, Korea Aerospace

University, Gyeonggi 412-791, Korea. (4)Department of Food and Nutrition, Seoul

National University, Seoul 151-742, Korea.

The accuracy of dietary assessments has emerged as a major concern in nutritional

epidemiology and new dietary assessment tools using computer technology to

increase accuracy have been developed in many countries. The purpose of this

study was to develop a web-based computer-assisted personal interview system

(CAPIS) for conducting dietary assessment and to evaluate its practical

utilization among Koreans. The client software was developed using Microsoft's

ClickOnce technology, which allows communication with a database system via an

http server to add or retrieve data. The system consists of a tracking system for

the subject and researcher, a data-input system during the interview, a

calculation system for estimating food and nutrient intake, a data-output system

for presenting the results, and an evaluation system for assessing the adequacy

of nutrient and food intake. Databases of the nutrient composition of common food

(n = 3,642), recipes for common dishes (n = 1,886), and photos of serving sizes

for food and dishes (n = 4,152) were constructed, and logical processes for data

collection, calculation, and output were developed. The functionality, on-site

applicability, and efficiency of CAPIS were evaluated in a convenience sample of

181 participants (61 males, 120 females; aged 24 to 85) by comparing with manual

24 hour recall method with paper questionnaire. The CAPIS was functioned

adequately in the field survey in terms of completeness of function, security,

and compliance of researcher and subjects. Regarding on-site applicability,

23.2%, 32.6%, 35.4%, and 43.7% of subjects reported that CAPIS was easier to

recall their diet, to estimate the amount consumed, to communicate with the

interviewer, and to concentrate on the interview than the manual method with

paper questionnaire, respectively. Although CAPIS required more interview time (9

min 42 sec) compared to the manual method (7 min 30 sec), it saved time and cost

for data coding and entry (15 min 35 sec) and gave high satisfaction from the

prompt feedback after interview to the subjects, which increase efficiency to

apply on the field survey. Our results suggest that the newly developed CAPIS is

suitable for conducting personal interviews for dietary assessment in Korean

population.

DOI: 10.7762/cnr.2014.3.2.115

PMCID: PMC4135239

PMID: 25136539

769. IEEE Trans Pattern Anal Mach Intell. 2014 Jul;36(7):1325-39. doi:

10.1109/TPAMI.2013.248.

Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in

Natural Environments.

Ionescu C, Papava D, Olaru V, Sminchisescu C.

We introduce a new dataset, Human3.6M, of 3.6 Million accurate 3D Human poses,

acquired by recording the performance of 5 female and 6 male subjects, under 4

different viewpoints, for training realistic human sensing systems and for

evaluating the next generation of human pose estimation models and algorithms.

Besides increasing the size of the datasets in the current state-of-the-art by

several orders of magnitude, we also aim to complement such datasets with a

diverse set of motions and poses encountered as part of typical human activities

(taking photos, talking on the phone, posing, greeting, eating, etc.), with

additional synchronized image, human motion capture, and time of flight (depth)

data, and with accurate 3D body scans of all the subject actors involved. We also

provide controlled mixed reality evaluation scenarios where 3D human models are

animated using motion capture and inserted using correct 3D geometry, in complex

real environments, viewed with moving cameras, and under occlusion. Finally, we

provide a set of large-scale statistical models and detailed evaluation baselines

for the dataset illustrating its diversity and the scope for improvement by

future work in the research community. Our experiments show that our best

large-scale model can leverage our full training set to obtain a 20% improvement

in performance compared to a training set of the scale of the largest existing

public dataset for this problem. Yet the potential for improvement by leveraging

higher capacity, more complex models with our large dataset, is substantially

vaster and should stimulate future research. The dataset together with code for

the associated large-scale learning models, features, visualization tools, as

well as the evaluation server, is available online at

http://vision.imar.ro/human3.6m.

DOI: 10.1109/TPAMI.2013.248

PMID: 26353306 [Indexed for MEDLINE]

770. J Appl Toxicol. 2014 Jul;34(7):805-9. doi: 10.1002/jat.2923. Epub 2013 Sep 11.

LTMap: a web server for assessing the potential liver toxicity by genome-wide

transcriptional expression data.

Xing L(1), Wu L, Liu Y, Ai N, Lu X, Fan X.

Author information:

(1)Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences,

Zhejiang University, Hangzhou, 310058, China.

Toxicogenomics (TGx) has played a significant role in mechanistic research

related with hepatotoxicity as well as liver toxicity prediction. Currently,

several large-scale preclinical TGx data sets were made freely accessible to the

public, such as Open TG-GATEs. With the availability of a sufficient amount of

microarray data, it is important to integrate this information to provide new

insights into the risk assessment of potential drug-induced liver toxicity. Here

we developed a web server for evaluating the potential liver toxicity based on

genome-wide transcriptomics data, namely LTMap. In LTMap, researchers could

compare signatures of query compounds against a pregenerated signature database

of 20 123 Affymetrix arrays associated with about 170 compounds retrieved from

the largest public toxicogenomics data set Open TG-GATEs. Results from this

comparison may lead to the unexpected discovery of similar toxicological

responses between chemicals. We validated our computational approach for

similarity comparison using three example drugs. Our successful applications of

LTMap in these case studies demonstrated its utility in revealing the connection

of chemicals according to similar toxicological behaviors. Furthermore, a

user-friendly web interface is provided by LTMap to browse and search

toxicogenomics data (http://tcm.zju.edu.cn/ltmap).

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/jat.2923

PMID: 24022982 [Indexed for MEDLINE]

771. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W356-60. doi:

10.1093/nar/gku459. Epub 2014 Jun 27.

SARA-Coffee web server, a tool for the computation of RNA sequence and structure

multiple alignments.

Di Tommaso P(1), Bussotti G(2), Kemena C(3), Capriotti E(4), Chatzou M(1), Prieto

P(1), Notredame C(5).

Author information:

(1)Comparative Bioinformatics, Bioinformatics and Genomics Program, Centre for

Genomic Regulation (CRG), Dr Aiguader 88, 08003 Barcelona, Spain Universitat

Pompeu Fabra (UPF), 08003 Barcelona, Spain. (2)European Molecular Biology

Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome

Campus, Hinxton, Cambridge CB10 1SD, UK. (3)Evolutionary Bioinformatics Group,

Institute for Evolution and Biodiversity, University of Münster, Hüfferstraße 1,

48145 Münster, Germany. (4)Division of Informatics, Department of Pathology,

University of Alabama at Birmingham, 35249 Birmingham (AL), USA. (5)Comparative

Bioinformatics, Bioinformatics and Genomics Program, Centre for Genomic

Regulation (CRG), Dr Aiguader 88, 08003 Barcelona, Spain Universitat Pompeu Fabra

(UPF), 08003 Barcelona, Spain cedric.notredame@crg.eu.

This article introduces the SARA-Coffee web server; a service allowing the online

computation of 3D structure based multiple RNA sequence alignments. The server

makes it possible to combine sequences with and without known 3D structures.

Given a set of sequences SARA-Coffee outputs a multiple sequence alignment along

with a reliability index for every sequence, column and aligned residue.

SARA-Coffee combines SARA, a pairwise structural RNA aligner with the R-Coffee

multiple RNA aligner in a way that has been shown to improve alignment accuracy

over most sequence aligners when enough structural data is available. The server

can be accessed from http://tcoffee.crg.cat/apps/tcoffee/do:saracoffee.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku459

PMCID: PMC4086076

PMID: 24972831 [Indexed for MEDLINE]

772. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W430-5. doi: 10.1093/nar/gku450.

Epub 2014 Jun 23.

PubServer: literature searches by homology.

Jaroszewski L(1), Koska L(2), Sedova M(2), Godzik A(3).

Author information:

(1)Joint Center for Structural Genomics (http://www.jcsg.org) Bioinformatics and

Systems Biology Program, Sanford Burnham Medical Research Institute, 10901 N.

Torrey Pines Road, La Jolla, CA 92037, USA Center for Research in Biological

Systems, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA

92093-0446, USA. (2)Joint Center for Structural Genomics (http://www.jcsg.org)

Bioinformatics and Systems Biology Program, Sanford Burnham Medical Research

Institute, 10901 N. Torrey Pines Road, La Jolla, CA 92037, USA. (3)Joint Center

for Structural Genomics (http://www.jcsg.org) Bioinformatics and Systems Biology

Program, Sanford Burnham Medical Research Institute, 10901 N. Torrey Pines Road,

La Jolla, CA 92037, USA Center for Research in Biological Systems, University of

California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0446, USA

adam@godziklab.org.

PubServer, available at http://pubserver.burnham.org/, is a tool to automatically

collect, filter and analyze publications associated with groups of homologous

proteins. Protein entries in databases such as Entrez Protein database at NCBI

contain information about publications associated with a given protein. The scope

of these publications varies a lot: they include studies focused on biochemical

functions of individual proteins, but also reports from genome sequencing

projects that introduce tens of thousands of proteins. Collecting and analyzing

publications related to sets of homologous proteins help in functional annotation

of novel protein families and in improving annotations of well-studied protein

families or individual genes. However, performing such collection and analysis

manually is a tedious and time-consuming process. PubServer automatically

collects identifiers of homologous proteins using PSI-Blast, retrieves literature

references from corresponding database entries and filters out publications

unlikely to contain useful information about individual proteins. It also

prepares simple vocabulary statistics from titles, abstracts and MeSH terms to

identify the most frequently occurring keywords, which may help to quickly

identify common themes in these publications. The filtering criteria applied to

collected publications are user-adjustable. The results of the server are

presented as an interactive page that allows re-filtering and different

presentations of the output.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku450

PMCID: PMC4086066

PMID: 24957597 [Indexed for MEDLINE]

773. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W331-6. doi: 10.1093/nar/gku483.

Epub 2014 Jun 6.

VarMod: modelling the functional effects of non-synonymous variants.

Pappalardo M(1), Wass MN(2).

Author information:

(1)Centre for Molecular Processing, School of Biosciences, University of Kent,

CT2 7NH, UK. (2)Centre for Molecular Processing, School of Biosciences,

University of Kent, CT2 7NH, UK m.n.wass@kent.ac.uk.

Unravelling the genotype-phenotype relationship in humans remains a challenging

task in genomics studies. Recent advances in sequencing technologies mean there

are now thousands of sequenced human genomes, revealing millions of single

nucleotide variants (SNVs). For non-synonymous SNVs present in proteins the

difficulties of the problem lie in first identifying those nsSNVs that result in

a functional change in the protein among the many non-functional variants and in

turn linking this functional change to phenotype. Here we present VarMod (Variant

Modeller) a method that utilises both protein sequence and structural features to

predict nsSNVs that alter protein function. VarMod develops recent observations

that functional nsSNVs are enriched at protein-protein interfaces and

protein-ligand binding sites and uses these characteristics to make predictions.

In benchmarking on a set of nearly 3000 nsSNVs VarMod performance is comparable

to an existing state of the art method. The VarMod web server provides extensive

resources to investigate the sequence and structural features associated with the

predictions including visualisation of protein models and complexes via an

interactive JSmol molecular viewer. VarMod is available for use at

http://www.wasslab.org/varmod.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku483

PMCID: PMC4086131

PMID: 24906884 [Indexed for MEDLINE]

774. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W137-46. doi:

10.1093/nar/gku412. Epub 2014 Jun 3.

DiseaseConnect: a comprehensive web server for mechanism-based disease-disease

connections.

Liu CC(1), Tseng YT(2), Li W(3), Wu CY(2), Mayzus I(4), Rzhetsky A(4), Sun F(3),

Waterman M(3), Chen JJ(5), Chaudhary PM(6), Loscalzo J(7), Crandall E(8), Zhou

XJ(9).

Author information:

(1)Institute of Genomics and Bioinformatics, National Chung Hsing University,

Taichung 402, Taiwan Agricultural Biotechnology Center, National Chung Hsing

University, Taichung 402, Taiwan xjzhou@usc.edu. (2)Institute of Genomics and

Bioinformatics, National Chung Hsing University, Taichung 402, Taiwan.

(3)Molecular and Computational Biology, University of Southern California, Los

Angeles, CA 90089, USA. (4)Computation Institute, Institute for Genomics and

Systems Biology University of Chicago, Chicago, IL 60637, USA. (5)Agricultural

Biotechnology Center, National Chung Hsing University, Taichung 402, Taiwan

Institute of Molecular Biology, National Chung Hsing University, Taichung 402,

Taiwan Institute of Biomedical Sciences, National Chung Hsing University,

Taichung 402, Taiwan. (6)Jane Anne Nohl Division of Hematology and Center for the

Study of Blood Diseases, University of Southern California Keck School of

Medicine, Los Angeles, CA 90033, USA. (7)Department of Medicine, Brigham and

Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. (8)Will

Rogers Institute Pulmonary Research Center, University of Southern California,

Los Angeles, CA 90033, USA Department of Medicine, University of Southern

California, Los Angeles, CA 90089, USA. (9)Molecular and Computational Biology,

University of Southern California, Los Angeles, CA 90089, USA xjzhou@usc.edu.

The DiseaseConnect (http://disease-connect.org) is a web server for analysis and

visualization of a comprehensive knowledge on mechanism-based disease

connectivity. The traditional disease classification system groups diseases with

similar clinical symptoms and phenotypic traits. Thus, diseases with entirely

different pathologies could be grouped together, leading to a similar treatment

design. Such problems could be avoided if diseases were classified based on their

molecular mechanisms. Connecting diseases with similar pathological mechanisms

could inspire novel strategies on the effective repositioning of existing drugs

and therapies. Although there have been several studies attempting to generate

disease connectivity networks, they have not yet utilized the enormous and

rapidly growing public repositories of disease-related omics data and literature,

two primary resources capable of providing insights into disease connections at

an unprecedented level of detail. Our DiseaseConnect, the first public web

server, integrates comprehensive omics and literature data, including a large

amount of gene expression data, Genome-Wide Association Studies catalog, and

text-mined knowledge, to discover disease-disease connectivity via common

molecular mechanisms. Moreover, the clinical comorbidity data and a comprehensive

compilation of known drug-disease relationships are additionally utilized for

advancing the understanding of the disease landscape and for facilitating the

mechanism-based development of new drug treatments.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku412

PMCID: PMC4086092

PMID: 24895436 [Indexed for MEDLINE]

775. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W94-9. doi: 10.1093/nar/gku436.

Epub 2014 Jun 3.

CFM-ID: a web server for annotation, spectrum prediction and metabolite

identification from tandem mass spectra.

Allen F(1), Pon A(2), Wilson M(2), Greiner R(2), Wishart D(2).

Author information:

(1)Department of Computing Science, Athabasca Hall, University of Alberta,

Edmonton T6G 2E8, Canada. felicity.allen@ualberta.ca. (2)Department of Computing

Science, Athabasca Hall, University of Alberta, Edmonton T6G 2E8, Canada.

CFM-ID is a web server supporting three tasks associated with the interpretation

of tandem mass spectra (MS/MS) for the purpose of automated metabolite

identification: annotation of the peaks in a spectrum for a known chemical

structure; prediction of spectra for a given chemical structure and putative

metabolite identification--a predicted ranking of possible candidate structures

for a target spectrum. The algorithms used for these tasks are based on

Competitive Fragmentation Modeling (CFM), a recently introduced probabilistic

generative model for the MS/MS fragmentation process that uses machine learning

techniques to learn its parameters from data. These algorithms have been

extensively tested on multiple datasets and have been shown to out-perform

existing methods such as MetFrag and FingerId. This web server provides a simple

interface for using these algorithms and a graphical display of the resulting

annotations, spectra and structures. CFM-ID is made freely available at

http://cfmid.wishartlab.com.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku436

PMCID: PMC4086103

PMID: 24895432 [Indexed for MEDLINE]

776. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W26-31. doi: 10.1093/nar/gku477.

Epub 2014 May 30.

SuperPred: update on drug classification and target prediction.

Nickel J(1), Gohlke BO(2), Erehman J(2), Banerjee P(3), Rong WW(4), Goede A(4),

Dunkel M(4), Preissner R(5).

Author information:

(1)Charité-University Medicine Berlin, Structural Bioinformatics Group, Institute

of Physiology & Experimental Clinical Research Center, Berlin 13125, Germany

Charité-University Medicine Berlin, Division of General Pediatrics, Department of

Pediatric Oncology and Hematology, Berlin 13353, Germany. (2)Charité-University

Medicine Berlin, Structural Bioinformatics Group, Institute of Physiology &

Experimental Clinical Research Center, Berlin 13125, Germany German Cancer

Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120,

Germany. (3)Charité-University Medicine Berlin, Structural Bioinformatics Group,

Institute of Physiology & Experimental Clinical Research Center, Berlin 13125,

Germany Graduate School of Computational System Biology, Berlin 10115, Germany.

(4)Charité-University Medicine Berlin, Structural Bioinformatics Group, Institute

of Physiology & Experimental Clinical Research Center, Berlin 13125, Germany.

(5)Charité-University Medicine Berlin, Structural Bioinformatics Group, Institute

of Physiology & Experimental Clinical Research Center, Berlin 13125, Germany

robert.preissner@charite.de.

The SuperPred web server connects chemical similarity of drug-like compounds with

molecular targets and the therapeutic approach based on the similar property

principle. Since the first release of this server, the number of known

compound-target interactions has increased from 7000 to 665,000, which allows not

only a better prediction quality but also the estimation of a confidence. Apart

from the addition of quantitative binding data and the statistical consideration

of the similarity distribution in all drug classes, new approaches were

implemented to improve the target prediction. The 3D similarity as well as the

occurrence of fragments and the concordance of physico-chemical properties is

also taken into account. In addition, the effect of different fingerprints on the

prediction was examined. The retrospective prediction of a drug class (ATC code

of the WHO) allows the evaluation of methods and descriptors for a

well-characterized set of approved drugs. The prediction is improved by 7.5% to a

total accuracy of 75.1%. For query compounds with sufficient structural

similarity, the web server allows prognoses about the medical indication area of

novel compounds and to find new leads for known targets. SuperPred is publicly

available without registration at: http://prediction.charite.de.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku477

PMCID: PMC4086135

PMID: 24878925 [Indexed for MEDLINE]

777. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W198-204. doi:

10.1093/nar/gku427. Epub 2014 May 30.

R. S. WebTool, a web server for random sampling-based significance evaluation of

pairwise distances.

Villiers F(1), Bastien O(2), Kwak JM(3).

Author information:

(1)Department of Cell Biology and Molecular Genetics, University of Maryland,

College Park, MD 20740, USA villiers@umd.edu. (2)Laboratoire de Physiologie

Cellulaire et Végétale, iRTSV, CEA-Grenoble, 17 rue des Martyrs, 38054 Grenoble

Cedex 9, France villiers@umd.edu. (3)Center for Plant Aging Research, Institute

for Basic Science, Department of New Biology, DGIST, Daegu 711-873, Republic of

Korea.

Pairwise comparison of data vectors represents a large part of computational

biology, especially with the continuous increase in genome-wide approaches

yielding more information from more biological samples simultaneously. Gene

clustering for function prediction as well as analyses of signalling pathways and

the time-dependent dynamics of a system are common biological approaches that

often rely on large dataset comparison. Different metrics can be used to evaluate

the similarity between entities to be compared, such as correlation coefficients

and distances. While the latter offers a more flexible way of measuring potential

biological relationships between datasets, the significance of any given distance

is highly dependent on the dataset and cannot be easily determined. Monte Carlo

methods are robust approaches for evaluating the significance of distance values

by multiple random permutations of the dataset followed by distance calculation.

We have developed R. S. WebTool (http://rswebtool.kwaklab.org), a user-friendly

online server for random sampling-based evaluation of distance significances that

features an array of visualization and analysis tools to help

non-bioinformaticist users extract significant relationships from random noise in

distance-based dataset analyses.

© Crown copyright 2014.

DOI: 10.1093/nar/gku427

PMCID: PMC4086074

PMID: 24878919 [Indexed for MEDLINE]

778. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W46-52. doi: 10.1093/nar/gku433.

Epub 2014 May 29.

DDI-CPI, a server that predicts drug-drug interactions through implementing the

chemical-protein interactome.

Luo H(1), Zhang P(2), Huang H(3), Huang J(4), Kao E(5), Shi L(6), He L(3), Yang

L(7).

Author information:

(1)Bio-X Institutes, Shanghai Jiao Tong University, Shanghai 200030, China

University of Arkansas at Little Rock/University of Arkansas for Medical

Sciences, Little Rock, AR 72204, USA. (2)Healthcare Analytics Research Group, IBM

T.J. Watson Research Center, Yorktown Heights, NY 10598, USA. (3)Bio-X

Institutes, Shanghai Jiao Tong University, Shanghai 200030, China. (4)Department

of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and

Harvard School of Public Heath, Boston, MA 02215, USA. (5)Department of

Bioengineering, University of California at Berkeley, Berkeley, CA 94720, USA.

(6)School of Pharmacy, Fudan University, Shanghai 201203, China. (7)Bio-X

Institutes, Shanghai Jiao Tong University, Shanghai 200030, China

lun.yang@gmail.com.

Drug-drug interactions (DDIs) may cause serious side-effects that draw great

attention from both academia and industry. Since some DDIs are mediated by

unexpected drug-human protein interactions, it is reasonable to analyze the

chemical-protein interactome (CPI) profiles of the drugs to predict their DDIs.

Here we introduce the DDI-CPI server, which can make real-time DDI predictions

based only on molecular structure. When the user submits a molecule, the server

will dock user's molecule across 611 human proteins, generating a CPI profile

that can be used as a feature vector for the pre-constructed prediction model. It

can suggest potential DDIs between the user's molecule and our library of 2515

drug molecules. In cross-validation and independent validation, the server

achieved an AUC greater than 0.85. Additionally, by investigating the CPI

profiles of predicted DDI, users can explore the PK/PD proteins that might be

involved in a particular DDI. A 3D visualization of the drug-protein interaction

will be provided as well. The DDI-CPI is freely accessible at

http://cpi.bio-x.cn/ddi/.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku433

PMCID: PMC4086096

PMID: 24875476 [Indexed for MEDLINE]

779. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W130-6. doi: 10.1093/nar/gku471.

Epub 2014 May 29.

PTHGRN: unraveling post-translational hierarchical gene regulatory networks using

PPI, ChIP-seq and gene expression data.

Guan D(1), Shao J(1), Zhao Z(1), Wang P(2), Qin J(2), Deng Y(3), Boheler KR(4),

Wang J(5), Yan B(6).

Author information:

(1)Department of Biology, Hong Kong Baptist University, Kowloon, Hong Kong SAR,

China. (2)Department of Biochemistry and HKU-SIRI, The University of Hong Kong,

Hong Kong SAR, China. (3)Department of Internal Medicine and Biochemistry, Rush

University Medical Center, Chicago, Illinois 60612, USA. (4)Stem Cell &

Regenerative Medicine Consortium, LKS Faculty of Medicine and Department of

Physiology, The University of Hong Kong, Hong Kong SAR, China bohelerk@hku.hk.

(5)Department of Biochemistry and HKU-SIRI, The University of Hong Kong, Hong

Kong SAR, China Centre for Genomic Sciences, LKS Faculty of Medicine, The

University of Hong Kong, Hong Kong SAR, China junwen@uw.edu. (6)Department of

Biology, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China Stem Cell &

Regenerative Medicine Consortium, LKS Faculty of Medicine and Department of

Physiology, The University of Hong Kong, Hong Kong SAR, China

yanbinai6017@gmail.com.

Interactions among transcriptional factors (TFs), cofactors and other proteins or

enzymes can affect transcriptional regulatory capabilities of eukaryotic

organisms. Post-translational modifications (PTMs) cooperate with TFs and

epigenetic alterations to constitute a hierarchical complexity in transcriptional

gene regulation. While clearly implicated in biological processes, our

understanding of these complex regulatory mechanisms is still limited and

incomplete. Various online software have been proposed for uncovering

transcriptional and epigenetic regulatory networks, however, there is a lack of

effective web-based software capable of constructing underlying interactive

organizations between post-translational and transcriptional regulatory

components. Here, we present an open web server, post-translational hierarchical

gene regulatory network (PTHGRN) to unravel relationships among PTMs, TFs,

epigenetic modifications and gene expression. PTHGRN utilizes a graphical

Gaussian model with partial least squares regression-based methodology, and is

able to integrate protein-protein interactions, ChIP-seq and gene expression data

and to capture essential regulation features behind high-throughput data. The

server provides an integrative platform for users to analyze ready-to-use public

high-throughput Omics resources or upload their own data for systems biology

study. Users can choose various parameters in the method, build network

topologies of interests and dissect their associations with biological functions.

Application of the software to stem cell and breast cancer demonstrates that it

is an effective tool for understanding regulatory mechanisms in biological

complex systems. PTHGRN web server is publically available at web site

http://www.byanbioinfo.org/pthgrn.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku471

PMCID: PMC4086064

PMID: 24875471 [Indexed for MEDLINE]

780. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W478-84. doi:

10.1093/nar/gku377. Epub 2014 May 26.

Enhancing UCSF Chimera through web services.

Huang CC(1), Meng EC(1), Morris JH(1), Pettersen EF(1), Ferrin TE(2).

Author information:

(1)Resource for Biocomputing, Visualization, and Informatics, Department of

Pharmaceutical Chemistry, University of California San Francisco, San Francisco,

CA 94143, USA. (2)Resource for Biocomputing, Visualization, and Informatics,

Department of Pharmaceutical Chemistry, University of California San Francisco,

San Francisco, CA 94143, USA tef@cgl.ucsf.edu.

Integrating access to web services with desktop applications allows for an

expanded set of application features, including performing computationally

intensive tasks and convenient searches of databases. We describe how we have

enhanced UCSF Chimera (http://www.rbvi.ucsf.edu/chimera/), a program for the

interactive visualization and analysis of molecular structures and related data,

through the addition of several web services

(http://www.rbvi.ucsf.edu/chimera/docs/webservices.html). By streamlining access

to web services, including the entire job submission, monitoring and retrieval

process, Chimera makes it simpler for users to focus on their science projects

rather than data manipulation. Chimera uses Opal, a toolkit for wrapping

scientific applications as web services, to provide scalable and transparent

access to several popular software packages. We illustrate Chimera's use of web

services with an example workflow that interleaves use of these services with

interactive manipulation of molecular sequences and structures, and we provide an

example Python program to demonstrate how easily Opal-based web services can be

accessed from within an application. Web server availability:

http://webservices.rbvi.ucsf.edu/opal2/dashboard?command=serviceList.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku377

PMCID: PMC4086125

PMID: 24861624 [Indexed for MEDLINE]

781. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W215-20. doi:

10.1093/nar/gku460. Epub 2014 May 26.

ProBiS-ligands: a web server for prediction of ligands by examination of protein

binding sites.

Konc J(1), Janežič D(2).

Author information:

(1)National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

konc@cmm.ki.si. (2)University of Primorska, Faculty of Mathematics, Natural

Sciences and Information Technologies, Glagoljaška 8, 6000 Koper, Slovenia

konc@cmm.ki.si.

The ProBiS-ligands web server predicts binding of ligands to a protein structure.

Starting with a protein structure or binding site, ProBiS-ligands first

identifies template proteins in the Protein Data Bank that share similar binding

sites. Based on the superimpositions of the query protein and the similar binding

sites found, the server then transposes the ligand structures from those sites to

the query protein. Such ligand prediction supports many activities, e.g. drug

repurposing. The ProBiS-ligands web server, an extension of the ProBiS web

server, is open and free to all users at http://probis.cmm.ki.si/ligands.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku460

PMCID: PMC4086080

PMID: 24861616 [Indexed for MEDLINE]

782. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W100-6. doi: 10.1093/nar/gku478.

Epub 2014 May 26.

PIQMIe: a web server for semi-quantitative proteomics data management and

analysis.

Kuzniar A(1), Kanaar R(2).

Author information:

(1)Department of Genetics, Cancer Genomics Netherlands, Erasmus Medical Center,

PO Box 2040, 3000 CA Rotterdam, The Netherlands. (2)Department of Genetics,

Cancer Genomics Netherlands, Erasmus Medical Center, PO Box 2040, 3000 CA

Rotterdam, The Netherlands Department of Radiation Oncology, Erasmus Medical

Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands.

We present the Proteomics Identifications and Quantitations Data Management and

Integration Service or PIQMIe that aids in reliable and scalable data management,

analysis and visualization of semi-quantitative mass spectrometry based

proteomics experiments. PIQMIe readily integrates peptide and (non-redundant)

protein identifications and quantitations from multiple experiments with

additional biological information on the protein entries, and makes the linked

data available in the form of a light-weight relational database, which enables

dedicated data analyses (e.g. in R) and user-driven queries. Using the web

interface, users are presented with a concise summary of their proteomics

experiments in numerical and graphical forms, as well as with a searchable

protein grid and interactive visualization tools to aid in the rapid assessment

of the experiments and in the identification of proteins of interest. The web

server not only provides data access through a web interface but also supports

programmatic access through RESTful web service. The web server is available at

http://piqmie.semiqprot-emc.cloudlet.sara.nl or

http://www.bioinformatics.nl/piqmie. This website is free and open to all users

and there is no login requirement.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku478

PMCID: PMC4086067

PMID: 24861615 [Indexed for MEDLINE]

783. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W290-5. doi: 10.1093/nar/gku437.

Epub 2014 May 22.

PredHS: a web server for predicting protein-protein interaction hot spots by

using structural neighborhood properties.

Deng L(1), Zhang QC(2), Chen Z(3), Meng Y(3), Guan J(4), Zhou S(5).

Author information:

(1)School of Software, Central South University, Changsha 410075, China

Department of Computer Science and Technology, Tongji University, Shanghai

201804, China. (2)Department of Biochemistry and Molecular Biophysics and Center

for Computational Biology and Bioinformatics, Columbia University, New York

10032, USAShanghai Key Lab of Intelligent Information Processing and School of

Computer Science, Fudan University, Shanghai 200433, China. (3)School of

Software, Central South University, Changsha 410075, China. (4)Department of

Computer Science and Technology, Tongji University, Shanghai 201804, China

sgzhou@fudan.edu.cn. (5)Shanghai Key Lab of Intelligent Information Processing

and School of Computer Science, Fudan University, Shanghai 200433, China

sgzhou@fudan.edu.cn.

Identifying specific hot spot residues that contribute significantly to the

affinity and specificity of protein interactions is a problem of the utmost

importance. We present an interactive web server, PredHS, which is based on an

effective structure-based hot spot prediction method. The PredHS prediction

method integrates many novel structural and energetic features with two types of

structural neighborhoods (Euclidian and Voronoi), and combines random forest and

sequential backward elimination algorithms to select an optimal subset of

features. PredHS achieved the highest performance identifying hot spots compared

with other state-of-the-art methods, as benchmarked by using an independent

experimentally verified dataset. The input to PredHS is protein structures in the

PDB format with at least two chains that form interfaces. Users can visualize

their predictions in an interactive 3D viewer and download the results as text

files. PredHS is available at http://www.predhs.org.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku437

PMCID: PMC4086081

PMID: 24852252 [Indexed for MEDLINE]

784. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W154-60. doi:

10.1093/nar/gku455. Epub 2014 May 22.

StemCellNet: an interactive platform for network-oriented investigations in stem

cell biology.

Pinto JP(1), Reddy Kalathur RK(2), Machado RS(2), Xavier JM(2), Bragança J(3),

Futschik ME(4).

Author information:

(1)Centre for Molecular and Structural Biomedicine, CBME/IBB, LA, University of

Algarve, Faro, Algarve 8005-139, Portugal jppinto@ualg.pt. (2)Centre for

Molecular and Structural Biomedicine, CBME/IBB, LA, University of Algarve, Faro,

Algarve 8005-139, Portugal. (3)Centre for Molecular and Structural Biomedicine,

CBME/IBB, LA, University of Algarve, Faro, Algarve 8005-139, Portugal Department

of Biomedical Sciences and Medicine, University of Algarve, Faro, Algarve

8005-139, Portugal. (4)Centre for Molecular and Structural Biomedicine, CBME/IBB,

LA, University of Algarve, Faro, Algarve 8005-139, Portugal Centre of Marine

Science, University of Algarve, Faro, Algarve 8005-139, Portugal jppinto@ualg.pt.

Stem cells are characterized by their potential for self-renewal and their

capacity to differentiate into mature cells. These two key features emerge

through the interplay of various factors within complex molecular networks. To

provide researchers with a dedicated tool to investigate these networks, we have

developed StemCellNet, a versatile web server for interactive network analysis

and visualization. It rapidly generates focused networks based on a large

collection of physical and regulatory interactions identified in human and murine

stem cells. The StemCellNet web-interface has various easy-to-use tools for

selection and prioritization of network components, as well as for integration of

expression data provided by the user. As a unique feature, the networks generated

can be screened against a compendium of stemness-associated genes. StemCellNet

can also indicate novel candidate genes by evaluating their connectivity

patterns. Finally, an optional dataset of generic interactions, which provides

large coverage of the human and mouse proteome, extends the versatility of

StemCellNet to other biomedical research areas in which stem cells play important

roles, such as in degenerative diseases or cancer. The StemCellNet web server is

freely accessible at http://stemcellnet.sysbiolab.eu.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku455

PMCID: PMC4086070

PMID: 24852251 [Indexed for MEDLINE]

785. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W344-9. doi: 10.1093/nar/gku448.

Epub 2014 May 22.

pocketZebra: a web-server for automated selection and classification of

subfamily-specific binding sites by bioinformatic analysis of diverse protein

families.

Suplatov D(1), Kirilin E(1), Arbatsky M(1), Takhaveev V(1), Svedas V(2).

Author information:

(1)Lomonosov Moscow State University, Belozersky Institute of Physicochemical

Biology and Faculty of Bioengineering and Bioinformatics, Vorobjev hills 1-73,

Moscow 119991, Russia. (2)Lomonosov Moscow State University, Belozersky Institute

of Physicochemical Biology and Faculty of Bioengineering and Bioinformatics,

Vorobjev hills 1-73, Moscow 119991, Russia vytas@belozersky.msu.ru.

The new web-server pocketZebra implements the power of bioinformatics and

geometry-based structural approaches to identify and rank subfamily-specific

binding sites in proteins by functional significance, and select particular

positions in the structure that determine selective accommodation of ligands. A

new scoring function has been developed to annotate binding sites by the presence

of the subfamily-specific positions in diverse protein families. pocketZebra

web-server has multiple input modes to meet the needs of users with different

experience in bioinformatics. The server provides on-site visualization of the

results as well as off-line version of the output in annotated text format and as

PyMol sessions ready for structural analysis. pocketZebra can be used to study

structure-function relationship and regulation in large protein superfamilies,

classify functionally important binding sites and annotate proteins with unknown

function. The server can be used to engineer ligand-binding sites and allosteric

regulation of enzymes, or implemented in a drug discovery process to search for

potential molecular targets and novel selective inhibitors/effectors. The server,

documentation and examples are freely available at

http://biokinet.belozersky.msu.ru/pocketzebra and there are no login

requirements.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku448

PMCID: PMC4086101

PMID: 24852248 [Indexed for MEDLINE]

786. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W350-5. doi: 10.1093/nar/gku396.

Epub 2014 May 21.

LocTree3 prediction of localization.

Goldberg T(1), Hecht M(2), Hamp T(2), Karl T(2), Yachdav G(3), Ahmed N(2),

Altermann U(2), Angerer P(2), Ansorge S(2), Balasz K(2), Bernhofer M(2), Betz

A(2), Cizmadija L(2), Do KT(2), Gerke J(2), Greil R(2), Joerdens V(2), Hastreiter

M(2), Hembach K(2), Herzog M(2), Kalemanov M(2), Kluge M(2), Meier A(2), Nasir

H(2), Neumaier U(2), Prade V(2), Reeb J(2), Sorokoumov A(2), Troshani I(2),

Vorberg S(2), Waldraff S(2), Zierer J(2), Nielsen H(4), Rost B(5).

Author information:

(1)Department of Informatics, Bioinformatics-I12, TUM, 85748 Garching, Germany

TUM Graduate School, Center of Doctoral Studies in Informatics and its

Applications (CeDoSIA), 85748 Garching, Germany goldberg@rostlab.org.

(2)Department of Informatics, Bioinformatics-I12, TUM, 85748 Garching, Germany.

(3)Department of Informatics, Bioinformatics-I12, TUM, 85748 Garching, Germany

Biosof LLC, New York, NY 10001, USA. (4)Center for Biological Sequence Analysis,

Department of Systems Biology, DTU, 2800 Lyngby, Denmark. (5)Department of

Informatics, Bioinformatics-I12, TUM, 85748 Garching, Germany Biosof LLC, New

York, NY 10001, USA Institute for Advanced Study (TUM-IAS), 85748 Garching,

Germany New York Consortium on Membrane Protein Structure (NYCOMPS) & Department

of Biochemistry and Molecular Biophysics, Columbia University, New York, NY

10032, USA Institute for Food and Plant Sciences WZW - Weihenstephan, 85350

Freising, Germany.

The prediction of protein sub-cellular localization is an important step toward

elucidating protein function. For each query protein sequence, LocTree2 applies

machine learning (profile kernel SVM) to predict the native sub-cellular

localization in 18 classes for eukaryotes, in six for bacteria and in three for

archaea. The method outputs a score that reflects the reliability of each

prediction. LocTree2 has performed on par with or better than any other

state-of-the-art method. Here, we report the availability of LocTree3 as a public

web server. The server includes the machine learning-based LocTree2 and improves

over it through the addition of homology-based inference. Assessed on

sequence-unique data, LocTree3 reached an 18-state accuracy Q18=80±3% for

eukaryotes and a six-state accuracy Q6=89±4% for bacteria. The server accepts

submissions ranging from single protein sequences to entire proteomes. Response

time of the unloaded server is about 90 s for a 300-residue eukaryotic protein

and a few hours for an entire eukaryotic proteome not considering the generation

of the alignments. For over 1000 entirely sequenced organisms, the predictions

are directly available as downloads. The web server is available at

http://www.rostlab.org/services/loctree3.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku396

PMCID: PMC4086075

PMID: 24848019 [Indexed for MEDLINE]

787. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W301-7. doi: 10.1093/nar/gku399.

Epub 2014 May 21.

PASTA 2.0: an improved server for protein aggregation prediction.

Walsh I(1), Seno F(2), Tosatto SC(3), Trovato A(2).

Author information:

(1)Department of Biomedical Sciences, University of Padova, Padova I-35131,

Italy. (2)INFN, Padova Section, and Department of Physics and Astronomy 'G.

Galilei', University of Padova, Padova I-35121, Italy. (3)Department of

Biomedical Sciences, University of Padova, Padova I-35131, Italy

silvio.tosatto@unipd.it.

The formation of amyloid aggregates upon protein misfolding is related to several

devastating degenerative diseases. The propensities of different protein

sequences to aggregate into amyloids, how they are enhanced by pathogenic

mutations, the presence of aggregation hot spots stabilizing pathological

interactions, the establishing of cross-amyloid interactions between

co-aggregating proteins, all rely at the molecular level on the stability of the

amyloid cross-beta structure. Our redesigned server, PASTA 2.0, provides a

versatile platform where all of these different features can be easily predicted

on a genomic scale given input sequences. The server provides other pieces of

information, such as intrinsic disorder and secondary structure predictions, that

complement the aggregation data. The PASTA 2.0 energy function evaluates the

stability of putative cross-beta pairings between different sequence stretches.

It was re-derived on a larger dataset of globular protein domains. The resulting

algorithm was benchmarked on comprehensive peptide and protein test sets, leading

to improved, state-of-the-art results with more amyloid forming regions correctly

detected at high specificity. The PASTA 2.0 server can be accessed at

http://protein.bio.unipd.it/pasta2/.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku399

PMCID: PMC4086119

PMID: 24848016 [Indexed for MEDLINE]

788. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W72-5. doi: 10.1093/nar/gku442.

Epub 2014 May 21.

EvoCor: a platform for predicting functionally related genes using phylogenetic

and expression profiles.

Dittmar WJ(1), McIver L(2), Michalak P(2), Garner HR(2), Valdez G(3).

Author information:

(1)Virginia Tech Carilion School of Medicine, Roanoke, VA 24016, USA. (2)Virginia

Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24061, USA. (3)Virginia

Tech Carilion Research Institute, Roanoke, VA 24016, USA; Department of

Biological Sciences, Virginia Tech, Blacksburg, VA 24061, USA

gvaldez1@vtc.vt.edu.

The wealth of publicly available gene expression and genomic data provides unique

opportunities for computational inference to discover groups of genes that

function to control specific cellular processes. Such genes are likely to have

co-evolved and be expressed in the same tissues and cells. Unfortunately, the

expertise and computational resources required to compare tens of genomes and

gene expression data sets make this type of analysis difficult for the average

end-user. Here, we describe the implementation of a web server that predicts

genes involved in affecting specific cellular processes together with a gene of

interest. We termed the server 'EvoCor', to denote that it detects functional

relationships among genes through evolutionary analysis and gene expression

correlation. This web server integrates profiles of sequence divergence derived

by a Hidden Markov Model (HMM) and tissue-wide gene expression patterns to

determine putative functional linkages between pairs of genes. This server is

easy to use and freely available at http://pilot-hmm.vbi.vt.edu/.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku442

PMCID: PMC4086105

PMID: 24848012 [Indexed for MEDLINE]

789. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W259-63. doi:

10.1093/nar/gku294. Epub 2014 May 16.

The CAD-score web server: contact area-based comparison of structures and

interfaces of proteins, nucleic acids and their complexes.

Olechnovič K(1), Venclovas C(2).

Author information:

(1)Institute of Biotechnology, Vilnius University, Graičiūno 8, Vilnius LT-02241,

Lithuania Faculty of Mathematics and Informatics, Vilnius University, Naugarduko

24, Vilnius LT-03225, Lithuania. (2)Institute of Biotechnology, Vilnius

University, Graičiūno 8, Vilnius LT-02241, Lithuania venclovas@ibt.lt.

The Contact Area Difference score (CAD-score) web server provides a universal

framework to compute and analyze discrepancies between different 3D structures of

the same biological macromolecule or complex. The server accepts both

single-subunit and multi-subunit structures and can handle all the major types of

macromolecules (proteins, RNA, DNA and their complexes). It can perform numerical

comparison of both structures and interfaces. In addition to entire structures

and interfaces, the server can assess user-defined subsets. The CAD-score server

performs both global and local numerical evaluations of structural differences

between structures or interfaces. The results can be explored interactively using

sortable tables of global scores, profiles of local errors, superimposed contact

maps and 3D structure visualization. The web server could be used for tasks such

as comparison of models with the native (reference) structure, comparison of

X-ray structures of the same macromolecule obtained in different states (e.g.

with and without a bound ligand), analysis of nuclear magnetic resonance (NMR)

structural ensemble or structures obtained in the course of molecular dynamics

simulation. The web server is freely accessible at:

http://www.ibt.lt/bioinformatics/cad-score.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku294

PMCID: PMC4086110

PMID: 24838571 [Indexed for MEDLINE]

790. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W240-5. doi: 10.1093/nar/gku394.

Epub 2014 May 16.

POSA: a user-driven, interactive multiple protein structure alignment server.

Li Z(1), Natarajan P(1), Ye Y(2), Hrabe T(1), Godzik A(3).

Author information:

(1)Bioinformatics and Systems Biology, Sanford-Burnham Medical Research

Institute, La Jolla, CA 92037, USA. (2)School of Informatics and Computing,

Indiana University, Bloomington, IN 47405, USA. (3)Bioinformatics and Systems

Biology, Sanford-Burnham Medical Research Institute, La Jolla, CA 92037, USA

adam@godziklab.org.

POSA (Partial Order Structure Alignment), available at http://posa.godziklab.org,

is a server for multiple protein structure alignment introduced in 2005 (Ye,Y.

and Godzik,A. (2005) Multiple flexible structure alignment using partial order

graphs. Bioinformatics, 21, 2362-2369). It is free and open to all users, and

there is no login requirement, albeit there is an option to register and store

results in individual, password-protected directories. In the updated POSA server

described here, we introduce two significant improvements. First is an interface

allowing the user to provide additional information by defining segments that

anchor the alignment in one or more input structures. This interface allows users

to take advantage of their intuition and biological insights to improve the

alignment and guide it toward a biologically relevant solution. The second

improvement is an interactive visualization with options that allow the user to

view all superposed structures in one window (a typical solution for visualizing

results of multiple structure alignments) or view them individually in a series

of synchronized windows with extensive, user-controlled visualization options.

The user can rotate structure(s) in any of the windows and study similarities or

differences between structures clearly visible in individual windows.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku394

PMCID: PMC4086100

PMID: 24838569 [Indexed for MEDLINE]

791. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W59-63. doi: 10.1093/nar/gku395.

Epub 2014 May 16.

SEPPA 2.0--more refined server to predict spatial epitope considering species of

immune host and subcellular localization of protein antigen.

Qi T(1), Qiu T(2), Zhang Q(1), Tang K(3), Fan Y(1), Qiu J(1), Wu D(1), Zhang

W(1), Chen Y(1), Gao J(4), Zhu R(1), Cao Z(5).

Author information:

(1)School of Life Sciences and Technology, Tongji University, Shanghai 200092,

China. (2)School of Life Sciences and Technology, Tongji University, Shanghai

200092, China zwcao@tongji.edu.cn. (3)School of Life Sciences and Technology,

Tongji University, Shanghai 200092, China Institute for Advanced Study of

Translational Medicine, Tongji University, Shanghai 200092, China. (4)College of

Information Engineering, Shanghai Maritime University, Shanghai 201306, China.

(5)School of Life Sciences and Technology, Tongji University, Shanghai 200092,

China Shanghai Center for Bioinformation and Technology, 1278 Keyuan Road,

Shanghai 201203, China.

Spatial Epitope Prediction server for Protein Antigens (SEPPA) has received lots

of feedback since being published in 2009. In this improved version, relative ASA

preference of unit patch and consolidated amino acid index were added as further

classification parameters in addition to unit-triangle propensity and clustering

coefficient which were previously reported. Then logistic regression model was

adopted instead of the previous simple additive one. Most importantly,

subcellular localization of protein antigen and species of immune host were fully

taken account to improve prediction. The result shows that AUC of 0.745 (5-fold

cross-validation) is almost the baseline performance with no differentiation like

all the other tools. Specifying subcellular localization of protein antigen and

species of immune host will generally push the AUC up. Secretory protein

immunized to mouse can push AUC to 0.823. In this version, the false positive

rate has been largely decreased as well. As the first method which has considered

the subcellular localization of protein antigen and species of immune host, SEPPA

2.0 shows obvious advantages over the other popular servers like SEPPA, PEPITO,

DiscoTope-2, B-pred, Bpredictor and Epitopia in supporting more specific

biological needs. SEPPA 2.0 can be accessed at http://badd.tongji.edu.cn/seppa/.

Batch query is also supported.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku395

PMCID: PMC4086087

PMID: 24838566 [Indexed for MEDLINE]

792. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W39-45. doi: 10.1093/nar/gku337.

Epub 2014 May 16.

DINIES: drug-target interaction network inference engine based on supervised

analysis.

Yamanishi Y(1), Kotera M(2), Moriya Y(3), Sawada R(4), Kanehisa M(3), Goto S(5).

Author information:

(1)Division of System Cohort, Medical Institute of Bioregulation, Kyushu

University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan Institute for

Advanced Study, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581,

Japan. (2)Graduate School of Bioscience and Biotechnology, Tokyo Institute of

Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8550, Japan. (3)Bioinformatics

Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011,

Japan. (4)Division of System Cohort, Medical Institute of Bioregulation, Kyushu

University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan.

(5)Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji,

Kyoto 611-0011, Japan goto@kuicr.kyoto-u.ac.jp.

DINIES (drug-target interaction network inference engine based on supervised

analysis) is a web server for predicting unknown drug-target interaction networks

from various types of biological data (e.g. chemical structures, drug side

effects, amino acid sequences and protein domains) in the framework of supervised

network inference. The originality of DINIES lies in prediction with

state-of-the-art machine learning methods, in the integration of heterogeneous

biological data and in compatibility with the KEGG database. The DINIES server

accepts any 'profiles' or precalculated similarity matrices (or 'kernels') of

drugs and target proteins in tab-delimited file format. When a training data set

is submitted to learn a predictive model, users can select either known

interaction information in the KEGG DRUG database or their own interaction data.

The user can also select an algorithm for supervised network inference, select

various parameters in the method and specify weights for heterogeneous data

integration. The server can provide integrative analyses with useful components

in KEGG, such as biological pathways, functional hierarchy and human diseases.

DINIES (http://www.genome.jp/tools/dinies/) is publicly available as one of the

genome analysis tools in GenomeNet.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku337

PMCID: PMC4086078

PMID: 24838565 [Indexed for MEDLINE]

793. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W53-8. doi: 10.1093/nar/gku401.

Epub 2014 May 16.

ProTox: a web server for the in silico prediction of rodent oral toxicity.

Drwal MN(1), Banerjee P(2), Dunkel M(1), Wettig MR(1), Preissner R(3).

Author information:

(1)Structural Bioinformatics Group, Institute for Physiology, Charité -

University Medicine Berlin, Lindenberger Weg 80, 13125 Berlin, Germany.

(2)Structural Bioinformatics Group, Institute for Physiology, Charité -

University Medicine Berlin, Lindenberger Weg 80, 13125 Berlin, Germany Graduate

School of Computational Systems Biology, Humboldt-Universität zu Berlin,

Invalidenstraße 42, 10115 Berlin, Germany. (3)Structural Bioinformatics Group,

Institute for Physiology, Charité - University Medicine Berlin, Lindenberger Weg

80, 13125 Berlin, Germany German Cancer Consortium (DKTK), Im Neuenheimer Feld

280, 69120 Heidelberg, Germany robert.preissner@charite.de.

Animal trials are currently the major method for determining the possible toxic

effects of drug candidates and cosmetics. In silico prediction methods represent

an alternative approach and aim to rationalize the preclinical drug development,

thus enabling the reduction of the associated time, costs and animal experiments.

Here, we present ProTox, a web server for the prediction of rodent oral toxicity.

The prediction method is based on the analysis of the similarity of compounds

with known median lethal doses (LD50) and incorporates the identification of

toxic fragments, therefore representing a novel approach in toxicity prediction.

In addition, the web server includes an indication of possible toxicity targets

which is based on an in-house collection of protein-ligand-based pharmacophore

models ('toxicophores') for targets associated with adverse drug reactions. The

ProTox web server is open to all users and can be accessed without registration

at: http://tox.charite.de/tox. The only requirement for the prediction is the

two-dimensional structure of the input compounds. All ProTox methods have been

evaluated based on a diverse external validation set and displayed strong

performance (sensitivity, specificity and precision of 76, 95 and 75%,

respectively) and superiority over other toxicity prediction tools, indicating

their possible applicability for other compound classes.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku401

PMCID: PMC4086068

PMID: 24838562 [Indexed for MEDLINE]

794. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W377-81. doi:

10.1093/nar/gku429. Epub 2014 May 15.

RASS: a web server for RNA alignment in the joint sequence-structure space.

He G(1), Steppi A(2), Laborde J(2), Srivastava A(2), Zhao P(1), Zhang J(3).

Author information:

(1)Department of Computer Science, Florida State University, Tallahassee, FL

32306, USA. (2)Department of Statistics, Florida State University, Tallahassee,

FL 32306, USA. (3)Department of Statistics, Florida State University,

Tallahassee, FL 32306, USA Jinfeng@stat.fsu.edu.

Comparison of ribonucleic acid (RNA) molecules is important for revealing their

evolutionary relationships, predicting their functions and predicting their

structures. Many methods have been developed for comparing RNAs using either

sequence or three-dimensional (3D) structure (backbone geometry) information.

Sequences and 3D structures contain non-overlapping sets of information that both

determine RNA functions. When comparing RNA 3D structures, both types of

information need to be taken into account. However, few methods compare RNA

structures using both sequence and 3D structure information. Recently, we have

developed a new method based on elastic shape analysis (ESA) that compares RNA

molecules by combining both sequence and 3D structure information. ESA treats RNA

structures as 3D curves with sequence information encoded on additional

coordinates so that the alignment can be performed in the joint

sequence-structure space. The similarity between two RNA molecules is quantified

by a formal distance, geodesic distance. In this study, we implement a web server

for the method, called RASS, to make it publicly available to research community.

The web server is located at http://cloud.stat.fsu.edu/RASS/.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku429

PMCID: PMC4086137

PMID: 24831547 [Indexed for MEDLINE]

795. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W308-13. doi:

10.1093/nar/gku369. Epub 2014 May 15.

AIDA: ab initio domain assembly server.

Xu D(1), Jaroszewski L(2), Li Z(1), Godzik A(3).

Author information:

(1)Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research

Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA.

(2)Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research

Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA Center for

Research in Biological Systems, University of California, San Diego, 9500 Gilman

Dr., La Jolla, CA 92093-0446, USA. (3)Bioinformatics and Systems Biology Program,

Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La

Jolla, CA 92037, USA Center for Research in Biological Systems, University of

California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0446, USA Center of

Excellence in Genomic Medicine Research (CEGMR), King Fahad Medical Research

Center, King Abdulaziz University, P.O. Box 80216, Jeddah 21589, Kingdom of Saudi

Arabia adam@godziklab.org.

AIDA: ab initio domain assembly server, available at

http://ffas.burnham.org/AIDA/ is a tool that can identify domains in multi-domain

proteins and then predict their 3D structures and relative spatial arrangements.

The server is free and open to all users, and there is an option for a user to

provide an e-mail to get the link to result page. Domains are evolutionary

conserved and often functionally independent units in proteins. Most proteins,

especially eukaryotic ones, consist of multiple domains while at the same time,

most experimentally determined protein structures contain only one or two

domains. As a result, often structures of individual domains in multi-domain

proteins can be accurately predicted, but the mutual arrangement of different

domains remains unknown. To address this issue we have developed AIDA program,

which combines steps of identifying individual domains, predicting (separately)

their structures and assembling them into multiple domain complexes using an ab

initio folding potential to describe domain-domain interactions. AIDA server not

only supports the assembly of a large number of continuous domains, but also

allows the assembly of domains inserted into other domains. Users can also

provide distance restraints to guide the AIDA energy minimization.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku369

PMCID: PMC4086082

PMID: 24831546 [Indexed for MEDLINE]

796. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W382-8. doi: 10.1093/nar/gku438.

Epub 2014 May 15.

COGNAC: a web server for searching and annotating hydrogen-bonded base

interactions in RNA three-dimensional structures.

Firdaus-Raih M(1), Hamdani HY(2), Nadzirin N(2), Ramlan EI(3), Willett P(4),

Artymiuk PJ(5).

Author information:

(1)School of Biosciences and Biotechnology, Faculty of Science and Technology,

Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia Institute of Systems

Biology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia

firdaus@mfrlab.org. (2)School of Biosciences and Biotechnology, Faculty of

Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi,

Malaysia. (3)Department of Artificial Intelligence, Faculty of Computer Science

and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia.

(4)Information School, University of Sheffield, Western Bank, Sheffield S10 2TN,

UK. (5)Department of Molecular Biology and Biotechnology, Krebs Institute,

University of Sheffield, Western Bank, Sheffield S10 2TN, UK.

Hydrogen bonds are crucial factors that stabilize a complex ribonucleic acid

(RNA) molecule's three-dimensional (3D) structure. Minute conformational changes

can result in variations in the hydrogen bond interactions in a particular

structure. Furthermore, networks of hydrogen bonds, especially those found in

tight clusters, may be important elements in structure stabilization or function

and can therefore be regarded as potential tertiary motifs. In this paper, we

describe a graph theoretical algorithm implemented as a web server that is able

to search for unbroken networks of hydrogen-bonded base interactions and thus

provide an accounting of such interactions in RNA 3D structures. This server,

COGNAC (COnnection tables Graphs for Nucleic ACids), is also able to compare the

hydrogen bond networks between two structures and from such annotations enable

the mapping of atomic level differences that may have resulted from

conformational changes due to mutations or binding events. The COGNAC server can

be accessed at http://mfrlab.org/grafss/cognac.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku438

PMCID: PMC4086061

PMID: 24831543 [Indexed for MEDLINE]

797. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W314-9. doi: 10.1093/nar/gku411.

Epub 2014 May 14.

DUET: a server for predicting effects of mutations on protein stability using an

integrated computational approach.

Pires DE(1), Ascher DB(2), Blundell TL(3).

Author information:

(1)Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK

dpires@dcc.ufmg.br. (2)Department of Biochemistry, University of Cambridge,

Cambridge, CB2 1GA, UK ACRF Rational Drug Discovery Centre and Biota Structural

Biology Laboratory, St Vincents Institute of Medical Research, Fitzroy, VIC 3065,

Australia. (3)Department of Biochemistry, University of Cambridge, Cambridge, CB2

1GA, UK tlb20@cam.ac.uk.

Cancer genome and other sequencing initiatives are generating extensive data on

non-synonymous single nucleotide polymorphisms (nsSNPs) in human and other

genomes. In order to understand the impacts of nsSNPs on the structure and

function of the proteome, as well as to guide protein engineering, accurate in

silicomethodologies are required to study and predict their effects on protein

stability. Despite the diversity of available computational methods in the

literature, none has proven accurate and dependable on its own under all

scenarios where mutation analysis is required. Here we present DUET, a web server

for an integrated computational approach to study missense mutations in proteins.

DUET consolidates two complementary approaches (mCSM and SDM) in a consensus

prediction, obtained by combining the results of the separate methods in an

optimized predictor using Support Vector Machines (SVM). We demonstrate that the

proposed method improves overall accuracy of the predictions in comparison with

either method individually and performs as well as or better than similar

methods. The DUET web server is freely and openly available at

http://structure.bioc.cam.ac.uk/duet.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku411

PMCID: PMC4086143

PMID: 24829462 [Indexed for MEDLINE]

798. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W361-7. doi: 10.1093/nar/gku406.

Epub 2014 May 14.

RBPmap: a web server for mapping binding sites of RNA-binding proteins.

Paz I(1), Kosti I(1), Ares M Jr(2), Cline M(3), Mandel-Gutfreund Y(4).

Author information:

(1)Department of Biology, Technion - Israel Institute of Technology, Technion

City, Haifa 32000, Israel. (2)Department of Molecular, Cellular and Developmental

Biology, UCSC, Santa Cruz, CA, USA. (3)Center for Biomolecular Science &

Engineering, UCSC, Santa Cruz, CA, USA. (4)Department of Biology, Technion -

Israel Institute of Technology, Technion City, Haifa 32000, Israel

yaelmg@tx.technion.ac.il.

Regulation of gene expression is executed in many cases by RNA-binding proteins

(RBPs) that bind to mRNAs as well as to non-coding RNAs. RBPs recognize their RNA

target via specific binding sites on the RNA. Predicting the binding sites of

RBPs is known to be a major challenge. We present a new webserver, RBPmap, freely

accessible through the website http://rbpmap.technion.ac.il/ for accurate

prediction and mapping of RBP binding sites. RBPmap has been developed

specifically for mapping RBPs in human, mouse and Drosophila melanogaster

genomes, though it supports other organisms too. RBPmap enables the users to

select motifs from a large database of experimentally defined motifs. In

addition, users can provide any motif of interest, given as either a consensus or

a PSSM. The algorithm for mapping the motifs is based on a Weighted-Rank

approach, which considers the clustering propensity of the binding sites and the

overall tendency of regulatory regions to be conserved. In addition, RBPmap

incorporates a position-specific background model, designed uniquely for

different genomic regions, such as splice sites, 5' and 3' UTRs, non-coding RNA

and intergenic regions. RBPmap was tested on high-throughput RNA-binding

experiments and was proved to be highly accurate.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku406

PMCID: PMC4086114

PMID: 24829458 [Indexed for MEDLINE]

799. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W285-9. doi: 10.1093/nar/gku397.

Epub 2014 May 14.

PRISM: a web server and repository for prediction of protein-protein interactions

and modeling their 3D complexes.

Baspinar A(1), Cukuroglu E(1), Nussinov R(2), Keskin O(3), Gursoy A(3).

Author information:

(1)Center for Computational Biology and Bioinformatics and College of

Engineering, Koc University, 34450 Istanbul, Turkey. (2)National Cancer

Institute, Cancer and Inflammation Program, Frederick National Laboratory for

Cancer Research, Leidos Biomedical Research, Inc., National Cancer Institute,

Frederick, MD 21702, USA Sackler Institute of Molecular Medicine, Department of

Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv

University, Tel Aviv 69978, Israel. (3)Center for Computational Biology and

Bioinformatics and College of Engineering, Koc University, 34450 Istanbul, Turkey

agursoy@ku.edu.tr okeskin@ku.edu.tr.

The PRISM web server enables fast and accurate prediction of protein-protein

interactions (PPIs). The prediction algorithm is knowledge-based. It combines

structural similarity and accounts for evolutionary conservation in the template

interfaces. The predicted models are stored in its repository. Given two protein

structures, PRISM will provide a structural model of their complex if a matching

template interface is available. Users can download the complex structure,

retrieve the interface residues and visualize the complex model. The PRISM web

server is user friendly, free and open to all users at

http://cosbi.ku.edu.tr/prism.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku397

PMCID: PMC4086120

PMID: 24829450 [Indexed for MEDLINE]

800. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W7-11. doi: 10.1093/nar/gku398.

Epub 2014 May 14.

Spaced words and kmacs: fast alignment-free sequence comparison based on inexact

word matches.

Horwege S(1), Lindner S(2), Boden M(2), Hatje K(3), Kollmar M(3), Leimeister

CA(2), Morgenstern B(4).

Author information:

(1)University of Göttingen, Institute of Microbiology and Genetics, Department of

Bioinformatics, Goldschmidtstraße 1, 37073 Göttingen, Germany

Sebastian.Horwege@stud.uni-goettingen.de. (2)University of Göttingen, Institute

of Microbiology and Genetics, Department of Bioinformatics, Goldschmidtstraße 1,

37073 Göttingen, Germany. (3)Max-Planck-Institute for Biophysical Chemistry,

Department of NMR-based Structural Biology, Group Systems Biology of Motor

Proteins, Am Fassberg 11, 37077 Göttingen, Germany. (4)University of Göttingen,

Institute of Microbiology and Genetics, Department of Bioinformatics,

Goldschmidtstraße 1, 37073 Göttingen, Germany Université d'Évry Val d'Essonne,

Laboratoire Statistique et Génome, UMR CNRS 8071, USC INRA, 23 Boulevard de

France, 91037 Évry, France.

In this article, we present a user-friendly web interface for two alignment-free

sequence-comparison methods that we recently developed. Most alignment-free

methods rely on exact word matches to estimate pairwise similarities or distances

between the input sequences. By contrast, our new algorithms are based on inexact

word matches. The first of these approaches uses the relative frequencies of

so-called spaced words in the input sequences, i.e. words containing 'don't care'

or 'wildcard' symbols at certain pre-defined positions. Various distance measures

can then be defined on sequences based on their different spaced-word

composition. Our second approach defines the distance between two sequences by

estimating for each position in the first sequence the length of the longest

substring at this position that also occurs in the second sequence with up to k

mismatches. Both approaches take a set of deoxyribonucleic acid (DNA) or protein

sequences as input and return a matrix of pairwise distance values that can be

used as a starting point for clustering algorithms or distance-based phylogeny

reconstruction. The two alignment-free programmes are accessible through a web

interface at 'Göttingen Bioinformatics Compute Server (GOBICS)':

http://spaced.gobics.de http://kmacs.gobics.de and the source codes can be

downloaded.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku398

PMCID: PMC4086093

PMID: 24829447 [Indexed for MEDLINE]

801. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W76-82. doi: 10.1093/nar/gku367.

Epub 2014 May 9.

WormNet v3: a network-assisted hypothesis-generating server for Caenorhabditis

elegans.

Cho A(1), Shin J(1), Hwang S(2), Kim C(1), Shim H(1), Kim H(1), Kim H(1), Lee

I(3).

Author information:

(1)Department of Biotechnology, College of Life Science and Biotechnology, Yonsei

University, Seoul, Korea. (2)Department of Biotechnology, College of Life Science

and Biotechnology, Yonsei University, Seoul, Korea Center for Systems and

Synthetic Biology, University of Texas at Austin, Austin, TX, USA. (3)Department

of Biotechnology, College of Life Science and Biotechnology, Yonsei University,

Seoul, Korea insuklee@yonsei.ac.kr.

High-throughput experimental technologies gradually shift the paradigm of

biological research from hypothesis-validation toward hypothesis-generation

science. Translating diverse types of large-scale experimental data into testable

hypotheses, however, remains a daunting task. We previously demonstrated that

heterogeneous genomics data can be integrated into a single genome-scale gene

network with high prediction power for ribonucleic acid interference (RNAi)

phenotypes in Caenorhabditis elegans, a popular metazoan model in the study of

developmental biology, neurobiology and genetics. Here, we present WormNet

version 3 (v3), which is a new network-assisted hypothesis-generating server for

C. elegans. WormNet v3 includes major updates to the base gene network, which

substantially improved predictions of RNAi phenotypes. The server generates

various gene network-based hypotheses using three complementary network methods:

(i) a phenotype-centric approach to 'find new members for a pathway'; (ii) a

gene-centric approach to 'infer functions from network neighbors' and (iii) a

context-centric approach to 'find context-associated hub genes', which is a new

method to identify key genes that mediate physiology within a specific context.

For example, we demonstrated that the context-centric approach can be used to

identify potential molecular targets of toxic chemicals. WormNet v3 is freely

accessible at http://www.inetbio.org/wormnet.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku367

PMCID: PMC4086142

PMID: 24813450 [Indexed for MEDLINE]

802. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W3-6. doi: 10.1093/nar/gku400.

Epub 2014 May 9.

Alignment-Annotator web server: rendering and annotating sequence alignments.

Gille C(1), Fähling M(2), Weyand B(3), Wieland T(4), Gille A(4).

Author information:

(1)Department of Biochemistry, Charité, University Medicine Berlin, Charitéplatz

1, 10117 Berlin, Germany christoph.gille@charite.de. (2)Institute of Vegetative

Physiology, Charité, University Medicine Berlin, Charitéplatz 1, 10117 Berlin,

Germany. (3)Department of Plastic, Hand and Reconstructive Surgery, Hannover

Medical School, Carl-Neubergstraße 1, 30625 Hannover, Germany. (4)Mannheim

Medical Faculty, Institute of Experimental and Clinical Pharmacology and

Toxicology, Heidelberg University, Maybachstraße 14, 68169 Mannheim, Germany.

Alignment-Annotator is a novel web service designed to generate interactive views

of annotated nucleotide and amino acid sequence alignments (i) de novo and (ii)

embedded in other software. All computations are performed at server side.

Interactivity is implemented in HTML5, a language native to web browsers. The

alignment is initially displayed using default settings and can be modified with

the graphical user interfaces. For example, individual sequences can be reordered

or deleted using drag and drop, amino acid color code schemes can be applied and

annotations can be added. Annotations can be made manually or imported (BioDAS

servers, the UniProt, the Catalytic Site Atlas and the PDB). Some edits take

immediate effect while others require server interaction and may take a few

seconds to execute. The final alignment document can be downloaded as a

zip-archive containing the HTML files. Because of the use of HTML the resulting

interactive alignment can be viewed on any platform including Windows, Mac OS X,

Linux, Android and iOS in any standard web browser. Importantly, no plugins nor

Java are required and therefore Alignment-Anotator represents the first

interactive browser-based alignment visualization.AVAILABILITY:

http://www.bioinformatics.org/strap/aa/ and http://strap.charite.de/aa/.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku400

PMCID: PMC4086088

PMID: 24813445 [Indexed for MEDLINE]

803. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W373-6. doi: 10.1093/nar/gku292.

Epub 2014 May 7.

e-RNA: a collection of web servers for comparative RNA structure prediction and

visualisation.

Lai D(1), Meyer IM(2).

Author information:

(1)Centre for High-Throughput Biology, Department of Computer Science and

Department of Medical Genetics, University of British Columbia, Vancouver V6T

1Z4, Canada. (2)Centre for High-Throughput Biology, Department of Computer

Science and Department of Medical Genetics, University of British Columbia,

Vancouver V6T 1Z4, Canada irmtraud.meyer@cantab.net.

e-RNA offers a free and open-access collection of five published RNA sequence

analysis tools, each solving specific problems not readily addressed by other

available tools. Given multiple sequence alignments, Transat detects all

conserved helices, including those expected in a final structure, but also

transient, alternative and pseudo-knotted helices. RNA-Decoder uses unique

evolutionary models to detect conserved RNA secondary structure in alignments

which may be partly protein-coding. SimulFold simultaneously co-estimates the

potentially pseudo-knotted conserved structure, alignment and phylogenetic tree

for a set of homologous input sequences. CoFold predicts the minimum-free energy

structure for an input sequence while taking the effects of co-transcriptional

folding into account, thereby greatly improving the prediction accuracy for long

sequences. R-chie is a program to visualise RNA secondary structures as arc

diagrams, allowing for easy comparison and analysis of conserved base-pairs and

quantitative features. The web site server dispatches user jobs to a cluster,

where up to 100 jobs can be processed in parallel. Upon job completion, users can

retrieve their results via a bookmarked or emailed link. e-RNA is located at

http://www.e-rna.org.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku292

PMCID: PMC4086097

PMID: 24810851 [Indexed for MEDLINE]

804. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W114-8. doi: 10.1093/nar/gku376.

Epub 2014 May 6.

STarMir: a web server for prediction of microRNA binding sites.

Rennie W(1), Liu C(1), Carmack CS(1), Wolenc A(1), Kanoria S(1), Lu J(2), Long

D(1), Ding Y(3).

Author information:

(1)Wadsworth Center, New York State Department of Health, Center for Medical

Science, 150 New Scotland Avenue, Albany, NY 12208, USA. (2)Department of

Genetics and Yale Stem Cell Center, Yale University, New Haven, CT 06520, USA.

(3)Wadsworth Center, New York State Department of Health, Center for Medical

Science, 150 New Scotland Avenue, Albany, NY 12208, USA yding@wadsworth.org.

STarMir web server predicts microRNA (miRNA) binding sites on a target

ribonucleic acid (RNA). STarMir is an implementation of logistic prediction

models developed with miRNA binding data from crosslinking immunoprecipitation

(CLIP) studies (Liu,C., Mallick, B., Long, D., Rennie, W.A., Wolenc, A., Carmack,

C.S. and Ding, Y. (2013). CLIP-based prediction of mammalian microRNA binding

sites. Nucleic Acids Res., 41(14), e138). In both intra-dataset and inter-dataset

validations, the models showed major improvements over established algorithms in

predictions of both seed and seedless sites. General applicability of the models

was indicated by good performance in cross-species validations. The input data

for STarMir is processed by the web server to perform prediction of miRNA binding

sites, compute comprehensive sequence, thermodynamic and target structure

features and a logistic probability as a measure of confidence for each predicted

site. For each of seed and seedless sites and for all three regions of a mRNA (3'

UTR, CDS and 5' UTR), STarMir output includes the computed binding site features,

the logistic probability and a publication-quality diagram of the predicted

miRNA:target hybrid. The prediction results are available through both an

interactive viewer and downloadable text files. As an application module of the

Sfold RNA package (http://sfold.wadsworth.org), STarMir is freely available to

all at http://sfold.wadsworth.org/starmir.html.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku376

PMCID: PMC4086099

PMID: 24803672 [Indexed for MEDLINE]

805. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W221-6. doi: 10.1093/nar/gku404.

Epub 2014 May 6.

PEP-SiteFinder: a tool for the blind identification of peptide binding sites on

protein surfaces.

Saladin A(1), Rey J(2), Thévenet P(1), Zacharias M(3), Moroy G(1), Tufféry P(4).

Author information:

(1)INSERM U973, MTi, F-75205 Paris, France Université Paris Diderot, Sorbonne

Paris Cité, F-75205 Paris, France. (2)INSERM U973, MTi, F-75205 Paris, France

Université Paris Diderot, Sorbonne Paris Cité, F-75205 Paris, France Ressource

Parisienne en Bioinformatique Structurale, F-75205 Paris, France. (3)Technische

Universität München 80333 München, Germany. (4)INSERM U973, MTi, F-75205 Paris,

France Université Paris Diderot, Sorbonne Paris Cité, F-75205 Paris, France

Ressource Parisienne en Bioinformatique Structurale, F-75205 Paris, France

pierre.tuffery@univ-paris-diderot.fr.

Peptide-protein interactions are important to many processes of life,

particularly for signal transmission or regulatory mechanisms. When no

information is known about the interaction between a protein and a peptide, it is

of interest to propose candidate sites of interaction at the protein surface, to

assist the design of biological experiments to probe the interaction, or to serve

as a starting point for more focused in silico approaches. PEP-SiteFinder is a

tool that will, given the structure of a protein and the sequence of a peptide,

identify protein residues predicted to be at peptide-protein interface.

PEP-SiteFinder relies on the 3D de novo generation of peptide conformations given

its sequence. These conformations then undergo a fast blind rigid docking on the

complete protein surface, and we have found, as the result of a benchmark over 41

complexes, that the best poses overlap to some extent the experimental patch of

interaction for close to 90% complexes. In addition, PEP-SiteFinder also returns

a propensity index we have found informative about the confidence of the

prediction. The PEP-SiteFinder web server is available at

http://bioserv.rpbs.univ-paris-diderot.fr/PEP-SiteFinder.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku404

PMCID: PMC4086095

PMID: 24803671 [Indexed for MEDLINE]

806. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W187-91. doi:

10.1093/nar/gku365. Epub 2014 May 5.

deepTools: a flexible platform for exploring deep-sequencing data.

Ramírez F(1), Dündar F(2), Diehl S(1), Grüning BA(3), Manke T(4).

Author information:

(1)Max Planck Institute of Immunobiology and Epigenetics, Stübeweg 51, 79108

Freiburg, Germany. (2)Max Planck Institute of Immunobiology and Epigenetics,

Stübeweg 51, 79108 Freiburg, Germany Faculty of Biology, University of Freiburg,

Schänzlestraße 1, 79104 Freiburg, Germany. (3)Department of Computer Science,

University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany. (4)Max

Planck Institute of Immunobiology and Epigenetics, Stübeweg 51, 79108 Freiburg,

Germany manke@ie-freiburg.mpg.de.

We present a Galaxy based web server for processing and visualizing deeply

sequenced data. The web server's core functionality consists of a suite of newly

developed tools, called deepTools, that enable users with little bioinformatic

background to explore the results of their sequencing experiments in a

standardized setting. Users can upload pre-processed files with continuous data

in standard formats and generate heatmaps and summary plots in a

straight-forward, yet highly customizable manner. In addition, we offer several

tools for the analysis of files containing aligned reads and enable efficient and

reproducible generation of normalized coverage files. As a modular and

open-source platform, deepTools can easily be expanded and customized to future

demands and developments. The deepTools webserver is freely available at

http://deeptools.ie-freiburg.mpg.de and is accompanied by extensive documentation

and tutorials aimed at conveying the principles of deep-sequencing data analysis.

The web server can be used without registration. deepTools can be installed

locally either stand-alone or as part of Galaxy.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku365

PMCID: PMC4086134

PMID: 24799436 [Indexed for MEDLINE]

807. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W337-43. doi:

10.1093/nar/gku366. Epub 2014 May 5.

PredictProtein--an open resource for online prediction of protein structural and

functional features.

Yachdav G(1), Kloppmann E(2), Kajan L(3), Hecht M(4), Goldberg T(4), Hamp T(3),

Hönigschmid P(5), Schafferhans A(3), Roos M(3), Bernhofer M(3), Richter L(3),

Ashkenazy H(6), Punta M(7), Schlessinger A(8), Bromberg Y(9), Schneider R(10),

Vriend G(11), Sander C(12), Ben-Tal N(13), Rost B(14).

Author information:

(1)Department of Informatics, Bioinformatics & Computational Biology i12, TUM

(Technische Universität München), Garching/Munich 85748, Germany Biosof LLC, New

York, NY 10001, USA TUM Graduate School, Center of Doctoral Studies in

Informatics and its Applications (CeDoSIA), TUM (Technische Universität München),

Garching/Munich 85748, Germany gyachdav@rostlab.org. (2)Department of

Informatics, Bioinformatics & Computational Biology i12, TUM (Technische

Universität München), Garching/Munich 85748, Germany New York Consortium on

Membrane Protein Structure (NYCOMPS), Columbia University, New York, NY 10032,

USA. (3)Department of Informatics, Bioinformatics & Computational Biology i12,

TUM (Technische Universität München), Garching/Munich 85748, Germany.

(4)Department of Informatics, Bioinformatics & Computational Biology i12, TUM

(Technische Universität München), Garching/Munich 85748, Germany TUM Graduate

School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA),

TUM (Technische Universität München), Garching/Munich 85748, Germany.

(5)Department of Genome Oriented Bioinformatics, Technische Universität München,

Wissenschaftszentrum Weihenstephan, Freising 85354, Germany. (6)The Department of

Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv

University, 69978 Tel Aviv, Israel. (7)Wellcome Trust Sanger Institute, Hinxton,

Cambridgeshire, CB10 1SA, UK Institute for Food and Plant Sciences

WZW-Weihenstephan, Alte Akademie 8, Freising 85350, Germany. (8)European

Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton,

Cambridgeshire, CB10 1SD, UK. (9)Biosof LLC, New York, NY 10001, USA Department

of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount

Sinai, One Gustave L. Levy Place, New York, NY 10029, USA. (10)Department of

Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, USA.

(11)Luxembourg University & Luxembourg Centre for Systems Biomedicine, 4362

Belval, Luxembourg. (12)CMBI, NCMLS, Radboudumc Nijmegen Medical Centre, 6525 GA

Nijmegen, The Netherlands. (13)Computational Biology Program, Memorial Sloan

Kettering Cancer Center, New York, 10065 NY, USA. (14)Department of Informatics,

Bioinformatics & Computational Biology i12, TUM (Technische Universität München),

Garching/Munich 85748, Germany Biosof LLC, New York, NY 10001, USA New York

Consortium on Membrane Protein Structure (NYCOMPS), Columbia University, New

York, NY 10032, USA The Department of Biochemistry and Molecular Biology, George

S. Wise Faculty of Life Sciences, Tel Aviv University, 69978 Tel Aviv, Israel

Department of Biochemistry and Molecular Biophysics & New York Consortium on

Membrane Protein Structure (NYCOMPS), Columbia University, New York, NY 10032,

USA Institute for Advanced Study (TUM-IAS), Garching/Munich 85748, Germany.

PredictProtein is a meta-service for sequence analysis that has been predicting

structural and functional features of proteins since 1992. Queried with a protein

sequence it returns: multiple sequence alignments, predicted aspects of structure

(secondary structure, solvent accessibility, transmembrane helices (TMSEG) and

strands, coiled-coil regions, disulfide bonds and disordered regions) and

function. The service incorporates analysis methods for the identification of

functional regions (ConSurf), homology-based inference of Gene Ontology terms

(metastudent), comprehensive subcellular localization prediction (LocTree3),

protein-protein binding sites (ISIS2), protein-polynucleotide binding sites

(SomeNA) and predictions of the effect of point mutations (non-synonymous SNPs)

on protein function (SNAP2). Our goal has always been to develop a system

optimized to meet the demands of experimentalists not highly experienced in

bioinformatics. To this end, the PredictProtein results are presented as both

text and a series of intuitive, interactive and visually appealing figures. The

web server and sources are available at http://ppopen.rostlab.org.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku366

PMCID: PMC4086098

PMID: 24799431 [Indexed for MEDLINE]

808. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W32-8. doi: 10.1093/nar/gku293.

Epub 2014 May 3.

SwissTargetPrediction: a web server for target prediction of bioactive small

molecules.

Gfeller D(1), Grosdidier A(1), Wirth M(1), Daina A(1), Michielin O(2), Zoete

V(3).

Author information:

(1)Swiss Institute of Bioinformatics (SIB), Quartier Sorge, Bâtiment Génopode,

CH-1015 Lausanne, Switzerland. (2)Swiss Institute of Bioinformatics (SIB),

Quartier Sorge, Bâtiment Génopode, CH-1015 Lausanne, Switzerland Ludwig Institute

for Cancer Research, Centre Hospitalier Universitaire Vaudois, Lausanne,

Switzerland Oncology Department, Centre Hospitalier Universitaire Vaudois,

Lausanne, Switzerland Olivier.Michielin@isb-sib.ch. (3)Swiss Institute of

Bioinformatics (SIB), Quartier Sorge, Bâtiment Génopode, CH-1015 Lausanne,

Switzerland Olivier.Michielin@isb-sib.ch.

Bioactive small molecules, such as drugs or metabolites, bind to proteins or

other macro-molecular targets to modulate their activity, which in turn results

in the observed phenotypic effects. For this reason, mapping the targets of

bioactive small molecules is a key step toward unraveling the molecular

mechanisms underlying their bioactivity and predicting potential side effects or

cross-reactivity. Recently, large datasets of protein-small molecule interactions

have become available, providing a unique source of information for the

development of knowledge-based approaches to computationally identify new targets

for uncharacterized molecules or secondary targets for known molecules. Here, we

introduce SwissTargetPrediction, a web server to accurately predict the targets

of bioactive molecules based on a combination of 2D and 3D similarity measures

with known ligands. Predictions can be carried out in five different organisms,

and mapping predictions by homology within and between different species is

enabled for close paralogs and orthologs. SwissTargetPrediction is accessible

free of charge and without login requirement at

http://www.swisstargetprediction.ch.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku293

PMCID: PMC4086140

PMID: 24792161 [Indexed for MEDLINE]

809. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W395-400. doi:

10.1093/nar/gku336. Epub 2014 Apr 29.

GeneGenie: optimized oligomer design for directed evolution.

Swainston N(1), Currin A(2), Day PJ(3), Kell DB(2).

Author information:

(1)Manchester Institute of Biotechnology, The University of Manchester,

Manchester M1 7DN, UK School of Computer Science, The University of Manchester,

Manchester M13 9PL, UK neil.swainston@manchester.ac.uk. (2)Manchester Institute

of Biotechnology, The University of Manchester, Manchester M1 7DN, UK School of

Chemistry, The University of Manchester, Manchester M13 9PL, UK. (3)Manchester

Institute of Biotechnology, The University of Manchester, Manchester M1 7DN, UK

Faculty of Medical and Human Sciences, The University of Manchester, Manchester

M13 9PT, UK.

GeneGenie, a new online tool available at http://www.gene-genie.org, is

introduced to support the design and self-assembly of synthetic genes and

constructs. GeneGenie allows for the design of oligonucleotide cohorts encoding

the gene sequence optimized for expression in any suitable host through an

intuitive, easy-to-use web interface. The tool ensures consistent oligomer

overlapping melting temperatures, minimizes the likelihood of misannealing,

optimizes codon usage for expression in a selected host, allows for specification

of forward and reverse cloning sequences (for downstream ligation) and also

provides support for mutagenesis or directed evolution studies. Directed

evolution studies are enabled through the construction of variant libraries via

the optional specification of 'variant codons', containing mixtures of bases, at

any position. For example, specifying the variant codon TNT (where N is any

nucleotide) will generate an equimolar mixture of the codons TAT, TCT, TGT and

TTT at that position, encoding a mixture of the amino acids Tyr, Ser, Cys and

Phe. This facility is demonstrated through the use of GeneGenie to develop and

synthesize a library of enhanced green fluorescent protein variants.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku336

PMCID: PMC4086129

PMID: 24782527 [Indexed for MEDLINE]

810. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W277-84. doi:

10.1093/nar/gku319. Epub 2014 Apr 29.

TSpred: a web server for the rational design of temperature-sensitive mutants.

Tan KP(1), Khare S(2), Varadarajan R(3), Madhusudhan MS(4).

Author information:

(1)Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore 138671

School of Computer Engineering, Nanyang Technological University, Singapore

639798. (2)Molecular Biophysics Unit, Indian Institute of Science, Bangalore

560012, India. (3)Molecular Biophysics Unit, Indian Institute of Science,

Bangalore 560012, India madhusudhan@iiserpune.ac.in. (4)Bioinformatics Institute,

30 Biopolis Street, #07-01, Matrix, Singapore 138671 Indian Institute of Science

Education and Research, Pune 411008, India Department of Biological Sciences,

National University of Singapore, Singapore 117543 madhusudhan@iiserpune.ac.in.

Temperature sensitive (Ts) mutants of proteins provide experimentalists with a

powerful and reversible way of conditionally expressing genes. The technique has

been widely used in determining the role of gene and gene products in several

cellular processes. Traditionally, Ts mutants are generated by random mutagenesis

and then selected though laborious large-scale screening. Our web server, TSpred

(http://mspc.bii.a-star.edu.sg/TSpred/), now enables users to rationally design

Ts mutants for their proteins of interest. TSpred uses hydrophobicity and

hydrophobic moment, deduced from primary sequence and residue depth, inferred

from 3D structures to predict/identify buried hydrophobic residues. Mutating

these residues leads to the creation of Ts mutants. Our method has been

experimentally validated in 36 positions in six different proteins. It is an

attractive proposition for Ts mutant engineering as it proposes a small number of

mutations and with high precision. The accompanying web server is simple and

intuitive to use and can handle proteins and protein complexes of different

sizes.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku319

PMCID: PMC4086094

PMID: 24782523 [Indexed for MEDLINE]

811. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W64-71. doi: 10.1093/nar/gku318.

Epub 2014 Apr 29.

NEP: web server for epitope prediction based on antibody neutralization of viral

strains with diverse sequences.

Chuang GY(1), Liou D(2), Kwong PD(1), Georgiev IS(3).

Author information:

(1)Vaccine Research Center. (2)Office of Cyber Infrastructure and Computational

Biology, National Institute of Allergy and Infectious Diseases, National

Institutes of Health, Bethesda, MD 20892, USA. (3)Vaccine Research Center

ivelin.georgiev@nih.gov.

Delineation of the antigenic site, or epitope, recognized by an antibody can

provide clues about functional vulnerabilities and resistance mechanisms, and can

therefore guide antibody optimization and epitope-based vaccine design.

Previously, we developed an algorithm for antibody-epitope prediction based on

antibody neutralization of viral strains with diverse sequences and validated the

algorithm on a set of broadly neutralizing HIV-1 antibodies. Here we describe the

implementation of this algorithm, NEP (Neutralization-based Epitope Prediction),

as a web-based server. The users must supply as input: (i) an alignment of

antigen sequences of diverse viral strains; (ii) neutralization data for the

antibody of interest against the same set of antigen sequences; and (iii)

(optional) a structure of the unbound antigen, for enhanced prediction accuracy.

The prediction results can be downloaded or viewed interactively on the antigen

structure (if supplied) from the web browser using a JSmol applet. Since

neutralization experiments are typically performed as one of the first steps in

the characterization of an antibody to determine its breadth and potency, the NEP

server can be used to predict antibody-epitope information at no additional

experimental costs. NEP can be accessed on the internet at

http://exon.niaid.nih.gov/nep.

Published by Oxford University Press on behalf of Nucleic Acids Research 2014.

This work is written by (a) US Government employee(s) and is in the public domain

in the US.

DOI: 10.1093/nar/gku318

PMCID: PMC4086065

PMID: 24782517 [Indexed for MEDLINE]

812. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W271-6. doi: 10.1093/nar/gku339.

Epub 2014 Apr 25.

iMODS: internal coordinates normal mode analysis server.

López-Blanco JR(1), Aliaga JI(2), Quintana-Ortí ES(2), Chacón P(3).

Author information:

(1)Department of Biological Chemical Physics, Rocasolano Physical Chemistry

Institute C.S.I.C., Serrano 119, 28006 Madrid, Spain. (2)Department of Computer

Science and Engineering, University Jaume I, 12071 Castellón, Spain.

(3)Department of Biological Chemical Physics, Rocasolano Physical Chemistry

Institute C.S.I.C., Serrano 119, 28006 Madrid, Spain pablo@chaconlab.org.

Normal mode analysis (NMA) in internal (dihedral) coordinates naturally

reproduces the collective functional motions of biological macromolecules. iMODS

facilitates the exploration of such modes and generates feasible transition

pathways between two homologous structures, even with large macromolecules. The

distinctive internal coordinate formulation improves the efficiency of NMA and

extends its applicability while implicitly maintaining stereochemistry.

Vibrational analysis, motion animations and morphing trajectories can be easily

carried out at different resolution scales almost interactively. The server is

versatile; non-specialists can rapidly characterize potential conformational

changes, whereas advanced users can customize the model resolution with multiple

coarse-grained atomic representations and elastic network potentials. iMODS

supports advanced visualization capabilities for illustrating collective motions,

including an improved affine-model-based arrow representation of domain dynamics.

The generated all-heavy-atoms conformations can be used to introduce flexibility

for more advanced modeling or sampling strategies. The server is free and open to

all users with no login requirement at http://imods.chaconlab.org.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku339

PMCID: PMC4086069

PMID: 24771341 [Indexed for MEDLINE]

813. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W210-4. doi: 10.1093/nar/gku321.

Epub 2014 Apr 21.

GalaxySite: ligand-binding-site prediction by using molecular docking.

Heo L(1), Shin WH(1), Lee MS(2), Seok C(3).

Author information:

(1)Department of Chemistry, Seoul National University, Seoul 151-747, Korea.

(2)Department of Biomedical Sciences, University of Ulsan College of Medicine,

Seoul 138-736, Korea chaok@snu.ac.kr. (3)Department of Chemistry, Seoul National

University, Seoul 151-747, Korea chaok@snu.ac.kr.

Knowledge of ligand-binding sites of proteins provides invaluable information for

functional studies, drug design and protein design. Recent progress in

ligand-binding-site prediction methods has demonstrated that using information

from similar proteins of known structures can improve predictions. The GalaxySite

web server, freely accessible at http://galaxy.seoklab.org/site, combines such

information with molecular docking for more precise binding-site prediction for

non-metal ligands. According to the recent critical assessments of structure

prediction methods held in 2010 and 2012, this server was found to be superior or

comparable to other state-of-the-art programs in the category of

ligand-binding-site prediction. A strong merit of the GalaxySite program is that

it provides additional predictions on binding ligands and their binding poses in

terms of the optimized 3D coordinates of the protein-ligand complexes, whereas

other methods predict only identities of binding-site residues or copy binding

geometry from similar proteins. The additional information on the specific

binding geometry would be very useful for applications in functional studies and

computer-aided drug discovery.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku321

PMCID: PMC4086128

PMID: 24753427 [Indexed for MEDLINE]

814. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W246-51. doi:

10.1093/nar/gku291. Epub 2014 Apr 21.

AlignMe--a membrane protein sequence alignment web server.

Stamm M(1), Staritzbichler R(1), Khafizov K(1), Forrest LR(2).

Author information:

(1)Computational Structural Biology Group, Max Planck Institute of Biophysics,

Frankfurt am Main 60438, Germany. (2)Computational Structural Biology Group, Max

Planck Institute of Biophysics, Frankfurt am Main 60438, Germany

lucy.forrest@nih.gov.

We present a web server for pair-wise alignment of membrane protein sequences,

using the program AlignMe. The server makes available two operational modes of

AlignMe: (i) sequence to sequence alignment, taking two sequences in fasta format

as input, combining information about each sequence from multiple sources and

producing a pair-wise alignment (PW mode); and (ii) alignment of two multiple

sequence alignments to create family-averaged hydropathy profile alignments (HP

mode). For the PW sequence alignment mode, four different optimized parameter

sets are provided, each suited to pairs of sequences with a specific similarity

level. These settings utilize different types of inputs: (position-specific)

substitution matrices, secondary structure predictions and transmembrane

propensities from transmembrane predictions or hydrophobicity scales. In the

second (HP) mode, each input multiple sequence alignment is converted into a

hydrophobicity profile averaged over the provided set of sequence homologs; the

two profiles are then aligned. The HP mode enables qualitative comparison of

transmembrane topologies (and therefore potentially of 3D folds) of two membrane

proteins, which can be useful if the proteins have low sequence similarity. In

summary, the AlignMe web server provides user-friendly access to a set of tools

for analysis and comparison of membrane protein sequences. Access is available at

http://www.bioinfo.mpg.de/AlignMe.

Published by Oxford University Press on behalf of Nucleic Acids Research 2014.

This work is written by (a) US Government employee(s) and is in the public domain

in the US.

DOI: 10.1093/nar/gku291

PMCID: PMC4086118

PMID: 24753425 [Indexed for MEDLINE]

815. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W124-9. doi: 10.1093/nar/gku317.

Epub 2014 Apr 21.

TargetRNA2: identifying targets of small regulatory RNAs in bacteria.

Kery MB(1), Feldman M(1), Livny J(2), Tjaden B(3).

Author information:

(1)Computer Science Department, Wellesley College, Wellesley, MA 02481, USA.

(2)Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. (3)Computer

Science Department, Wellesley College, Wellesley, MA 02481, USA

btjaden@wellesley.edu.

Many small, noncoding RNAs (sRNAs) in bacteria act as posttranscriptional

regulators of messenger RNAs. TargetRNA2 is a web server that identifies mRNA

targets of sRNA regulatory action in bacteria. As input, TargetRNA2 takes the

sequence of an sRNA and the name of a sequenced bacterial replicon. When

searching for targets of RNA regulation, TargetRNA2 uses a variety of features,

including conservation of the sRNA in other bacteria, the secondary structure of

the sRNA, the secondary structure of each candidate mRNA target and the

hybridization energy between the sRNA and each candidate mRNA target. TargetRNA2

outputs a ranked list of likely regulatory targets for the input sRNA. When

evaluated on a comprehensive set of sRNA-target interactions, TargetRNA2 was

found to be both accurate and efficient in identifying targets of sRNA regulatory

action. Furthermore, TargetRNA2 has the ability to integrate RNA-seq data, if

available. If an sRNA is differentially expressed in two or more RNA-seq

experiments, TargetRNA2 considers co-differential gene expression when searching

for regulatory targets, significantly improving the accuracy of target

identifications. The TargetRNA2 web server is freely available for use at

http://cs.wellesley.edu/∼btjaden/TargetRNA2.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku317

PMCID: PMC4086111

PMID: 24753424 [Indexed for MEDLINE]

816. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W320-4. doi: 10.1093/nar/gku316.

Epub 2014 Apr 21.

Deciphering key features in protein structures with the new ENDscript server.

Robert X(1), Gouet P(1).

Author information:

(1)Equipe de Biocristallographie et Biologie Structurale des Cibles

Thérapeutiques, IBCP-BMSSI, UMR 5086 CNRS Université de Lyon, SFR BioSciences

Gerland-Lyon Sud, 7 Passage du Vercors, 69367 Lyon Cedex 07, France

patrice.gouet@ibcp.fr.

ENDscript 2 is a friendly Web server for extracting and rendering a comprehensive

analysis of primary to quaternary protein structure information in an automated

way. This major upgrade has been fully re-engineered to enhance speed, accuracy

and usability with interactive 3D visualization. It takes advantage of the new

version 3 of ESPript, our well-known sequence alignment renderer, improved to

handle a large number of data with reduced computation time. From a single PDB

entry or file, ENDscript produces high quality figures displaying multiple

sequence alignment of proteins homologous to the query, colored according to

residue conservation. Furthermore, the experimental secondary structure elements

and a detailed set of relevant biophysical and structural data are depicted. All

this information and more are now mapped on interactive 3D PyMOL representations.

Thanks to its adaptive and rigorous algorithm, beginner to expert users can

modify settings to fine-tune ENDscript to their needs. ENDscript has also been

upgraded as an open platform for the visualization of multiple biochemical and

structural data coming from external biotool Web servers, with both 2D and 3D

representations. ENDscript 2 and ESPript 3 are freely available at

http://endscript.ibcp.fr and http://espript.ibcp.fr, respectively.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku316

PMCID: PMC4086106

PMID: 24753421 [Indexed for MEDLINE]

817. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W12-9. doi: 10.1093/nar/gku315.

Epub 2014 Apr 21.

DMINDA: an integrated web server for DNA motif identification and analyses.

Ma Q(1), Zhang H(2), Mao X(1), Zhou C(3), Liu B(3), Chen X(2), Xu Y(4).

Author information:

(1)Computational Systems Biology Laboratory, Department of Biochemistry and

Molecular Biology, and Institute of Bioinformatics, University of

Georgia, Athens GA 30602, USA BioEnergy Science Center (BESC), Oak Ridge National

Laboratory, Oak Ridge, Tennessee 37831, USA. (2)Computational Systems Biology

Laboratory, Department of Biochemistry and Molecular Biology, and Institute of

Bioinformatics, University of Georgia, Athens GA 30602, USA College of Computer

Science and Technology, Jilin University, Changchun, China. (3)School of

Mathematics, Shandong University, Jinan, Shandong, China. (4)Computational

Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, and

Institute of Bioinformatics, University of Georgia, Athens GA 30602, USA

BioEnergy Science Center (BESC), Oak Ridge National Laboratory, Oak Ridge,

Tennessee 37831, USA College of Computer Science and Technology, Jilin

University, Changchun, China xyn@bmb.uga.edu.

DMINDA (DNA motif identification and analyses) is an integrated web server for

DNA motif identification and analyses, which is accessible at

http://csbl.bmb.uga.edu/DMINDA/. This web site is freely available to all users

and there is no login requirement. This server provides a suite of cis-regulatory

motif analysis functions on DNA sequences, which are important to elucidation of

the mechanisms of transcriptional regulation: (i) de novo motif finding for a

given set of promoter sequences along with statistical scores for the predicted

motifs derived based on information extracted from a control set, (ii) scanning

motif instances of a query motif in provided genomic sequences, (iii) motif

comparison and clustering of identified motifs, and (iv) co-occurrence analyses

of query motifs in given promoter sequences. The server is powered by a backend

computer cluster with over 150 computing nodes, and is particularly useful for

motif prediction and analyses in prokaryotic genomes. We believe that DMINDA, as

a new and comprehensive web server for cis-regulatory motif finding and analyses,

will benefit the genomic research community in general and prokaryotic genome

researchers in particular.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic

Acids Research.

DOI: 10.1093/nar/gku315

PMCID: PMC4086085

PMID: 24753419 [Indexed for MEDLINE]

818. PLoS One. 2014 Jun 26;9(6):e100278. doi: 10.1371/journal.pone.0100278.

eCollection 2014.

Prediction of membrane transport proteins and their substrate specificities using

primary sequence information.

Mishra NK(1), Chang J(1), Zhao PX(1).

Author information:

(1)Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore,

Oklahoma, United States of America.

BACKGROUND: Membrane transport proteins (transporters) move hydrophilic

substrates across hydrophobic membranes and play vital roles in most cellular

functions. Transporters represent a diverse group of proteins that differ in

topology, energy coupling mechanism, and substrate specificity as well as

sequence similarity. Among the functional annotations of transporters,

information about their transporting substrates is especially important. The

experimental identification and characterization of transporters is currently

costly and time-consuming. The development of robust bioinformatics-based methods

for the prediction of membrane transport proteins and their substrate

specificities is therefore an important and urgent task.

RESULTS: Support vector machine (SVM)-based computational models, which

comprehensively utilize integrative protein sequence features such as amino acid

composition, dipeptide composition, physico-chemical composition, biochemical

composition, and position-specific scoring matrices (PSSM), were developed to

predict the substrate specificity of seven transporter classes: amino acid,

anion, cation, electron, protein/mRNA, sugar, and other transporters. An

additional model to differentiate transporters from non-transporters was also

developed. Among the developed models, the biochemical composition and PSSM

hybrid model outperformed other models and achieved an overall average prediction

accuracy of 76.69% with a Mathews correlation coefficient (MCC) of 0.49 and a

receiver operating characteristic area under the curve (AUC) of 0.833 on our main

dataset. This model also achieved an overall average prediction accuracy of

78.88% and MCC of 0.41 on an independent dataset.

CONCLUSIONS: Our analyses suggest that evolutionary information (i.e., the PSSM)

and the AAIndex are key features for the substrate specificity prediction of

transport proteins. In comparison, similarity-based methods such as BLAST,

PSI-BLAST, and hidden Markov models do not provide accurate predictions for the

substrate specificity of membrane transport proteins. TrSSP: The Transporter

Substrate Specificity Prediction Server, a web server that implements the SVM

models developed in this paper, is freely available at

http://bioinfo.noble.org/TrSSP.

DOI: 10.1371/journal.pone.0100278

PMCID: PMC4072671

PMID: 24968309 [Indexed for MEDLINE]

819. Int J Mol Sci. 2014 Jun 25;15(7):11204-19. doi: 10.3390/ijms150711204.

PSNO: predicting cysteine S-nitrosylation sites by incorporating various

sequence-derived features into the general form of Chou's PseAAC.

Zhang J(1), Zhao X(2), Sun P(3), Ma Z(4).

Author information:

(1)School of Computer Science and Information Technology, Northeast Normal

University, Changchun 130017, China. zhangj943@nenu.edu.cn. (2)School of Computer

Science and Information Technology, Northeast Normal University, Changchun

130017, China. xwzhao\_nenu@yeah.net. (3)School of Computer Science and

Information Technology, Northeast Normal University, Changchun 130017, China.

ppsun\_nenu@yeah.net. (4)School of Computer Science and Information Technology,

Northeast Normal University, Changchun 130017, China. zhiqiangma.nenu@gmail.com.

S-nitrosylation (SNO) is one of the most universal reversible post-translational

modifications involved in many biological processes. Malfunction or dysregulation

of SNO leads to a series of severe diseases, such as developmental abnormalities

and various diseases. Therefore, the identification of SNO sites (SNOs) provides

insights into disease progression and drug development. In this paper, a new

bioinformatics tool, named PSNO, is proposed to identify SNOs from protein

sequences. Firstly, we explore various promising sequence-derived discriminative

features, including the evolutionary profile, the predicted secondary structure

and the physicochemical properties. Secondly, rather than simply combining the

features, which may bring about information redundancy and unwanted noise, we use

the relative entropy selection and incremental feature selection approach to

select the optimal feature subsets. Thirdly, we train our model by the technique

of the k-nearest neighbor algorithm. Using both informative features and an

elaborate feature selection scheme, our method, PSNO, achieves good prediction

performance with a mean Mathews correlation coefficient (MCC) value of about

0.5119 on the training dataset using 10-fold cross-validation. These results

indicate that PSNO can be used as a competitive predictor among the

state-of-the-art SNOs prediction tools. A web-server, named PSNO, which

implements the proposed method, is freely available at

http://59.73.198.144:8088/PSNO/.

DOI: 10.3390/ijms150711204

PMCID: PMC4139777

PMID: 24968264 [Indexed for MEDLINE]

820. BMC Bioinformatics. 2014 Jun 24;15:214. doi: 10.1186/1471-2105-15-214.

e!DAL--a framework to store, share and publish research data.

Arend D(1), Lange M, Chen J, Colmsee C, Flemming S, Hecht D, Scholz U.

Author information:

(1)Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), OT

Gatersleben, Corrensstr, 3, 06466 Stadt Seeland, Germany.

arendd@ipk-gatersleben.de.

BACKGROUND: The life-science community faces a major challenge in handling "big

data", highlighting the need for high quality infrastructures capable of sharing

and publishing research data. Data preservation, analysis, and publication are

the three pillars in the "big data life cycle". The infrastructures currently

available for managing and publishing data are often designed to meet

domain-specific or project-specific requirements, resulting in the repeated

development of proprietary solutions and lower quality data publication and

preservation overall.

RESULTS: e!DAL is a lightweight software framework for publishing and sharing

research data. Its main features are version tracking, metadata management,

information retrieval, registration of persistent identifiers (DOI), an embedded

HTTP(S) server for public data access, access as a network file system, and a

scalable storage backend. e!DAL is available as an API for local non-shared

storage and as a remote API featuring distributed applications. It can be

deployed "out-of-the-box" as an on-site repository.

CONCLUSIONS: e!DAL was developed based on experiences coming from decades of

research data management at the Leibniz Institute of Plant Genetics and Crop

Plant Research (IPK). Initially developed as a data publication and documentation

infrastructure for the IPK's role as a data center in the DataCite consortium,

e!DAL has grown towards being a general data archiving and publication

infrastructure. The e!DAL software has been deployed into the Maven Central

Repository. Documentation and Software are also available at:

http://edal.ipk-gatersleben.de.

DOI: 10.1186/1471-2105-15-214

PMCID: PMC4080583

PMID: 24958009 [Indexed for MEDLINE]

821. Genet Mol Res. 2014 Jun 17;13(2):4564-72. doi: 10.4238/2014.June.17.8.

Predicting bacterial essential genes using only sequence composition information.

Ning LW(1), Lin H(1), Ding H(1), Huang J(1), Rao N(1), Guo FB(2).

Author information:

(1)Center of Bioinformatics and Key Laboratory for NeuroInformation of Ministry

of Education, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu, China. (2)Center of Bioinformatics and

Key Laboratory for NeuroInformation of Ministry of Education, School of Life

Science and Technology, University of Electronic Science and Technology of China,

Chengdu, China fbguo@uestc.edu.cn.

Essential genes are those genes that are needed by organisms at any time and

under any conditions. It is very important for us to identify essential genes

from bacterial genomes because of their vital role in synthetic biology and

biomedical practices. In this paper, we developed a support vector machine

(SVM)-based method to predict essential genes of bacterial genomes using only

compositional features. These features are all derived from the primary

sequences, i.e., nucleotide sequences and protein sequences. After training on

the multiple samplings of the labeled (essential or not essential) features using

a library for SVM, we obtained an average area under the ROC curve (AUC) of about

0.82 in a 5-fold cross-validation for Escherichia coli and about 0.74 for

Mycoplasma pulmonis. We further evaluated the performance of the method proposed

using the dataset consisting of 16 bacterial genomes, and an average AUC of 0.76

was achieved. Based on this training dataset, a model for essential gene

prediction was established. Another two independent genomes, Shewanella

oneidensis RW1 and Salmonella enterica serovar Typhimurium SL1344 were used to

evalutate the model. Results showed that the AUC sores were 0.77 and 0.81,

respectively. For the convenience of the vast majority of experimental

scientists, a web server has been constructed, which is freely available at

http://cefg.uestc.edu.cn:9999/egp.

DOI: 10.4238/2014.June.17.8

PMID: 25036505 [Indexed for MEDLINE]

822. Bioinformatics. 2014 Jun 15;30(12):1791-2. doi: 10.1093/bioinformatics/btu103.

Epub 2014 Feb 25.

Canto: an online tool for community literature curation.

Rutherford KM(1), Harris MA(1), Lock A(2), Oliver SG(1), Wood V(1).

Author information:

(1)Cambridge Systems Biology Centre, Department of Biochemistry, University of

Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1GA and

Department of Genetics, Evolution and Environment, and UCL Cancer Institute,

University College London, Darwin Building, Gower Street, London WC1E 6BT,

UKCambridge Systems Biology Centre, Department of Biochemistry, University of

Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1GA and

Department of Genetics, Evolution and Environment, and UCL Cancer Institute,

University College London, Darwin Building, Gower Street, London WC1E 6BT, UK.

(2)Cambridge Systems Biology Centre, Department of Biochemistry, University of

Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1GA and

Department of Genetics, Evolution and Environment, and UCL Cancer Institute,

University College London, Darwin Building, Gower Street, London WC1E 6BT, UK.

MOTIVATION: Detailed curation of published molecular data is essential for any

model organism database. Community curation enables researchers to contribute

data from their papers directly to databases, supplementing the activity of

professional curators and improving coverage of a growing body of literature. We

have developed Canto, a web-based tool that provides an intuitive curation

interface for both curators and researchers, to support community curation in the

fission yeast database, PomBase. Canto supports curation using OBO ontologies,

and can be easily configured for use with any species.

AVAILABILITY: Canto code and documentation are available under an Open Source

license from http://curation.pombase.org/. Canto is a component of the Generic

Model Organism Database (GMOD) project (http://www.gmod.org/).

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu103

PMCID: PMC4058955

PMID: 24574118 [Indexed for MEDLINE]

823. Bioinformatics. 2014 Jun 15;30(12):1780-1. doi: 10.1093/bioinformatics/btu109.

Epub 2014 Feb 25.

MGDB: crossing the marker genes of a user microarray with a database of

public-microarrays marker genes.

Huerta M(1), Munyi M(1), Expósito D(1), Querol E(1), Cedano J(1).

Author information:

(1)Institut de Biotecnologia i Biomedicina, Escola Tècnica Superior de

Ingenieria, Universitat Autònoma de Barcelona; Bellaterra, Barcelona 08193, Spain

and Laboratorio de Inmunología, Universidad de la República Regional Norte-Salto,

Rivera 1350, CP 50000 Salto, Uruguay.

SUMMARY: The microarrays performed by scientific teams grow exponentially. These

microarray data could be useful for researchers around the world, but

unfortunately they are underused. To fully exploit these data, it is necessary

(i) to extract these data from a repository of the high-throughput gene

expression data like Gene Expression Omnibus (GEO) and (ii) to make the data from

different microarrays comparable with tools easy to use for scientists. We have

developed these two solutions in our server, implementing a database of

microarray marker genes (Marker Genes Data Base). This database contains the

marker genes of all GEO microarray datasets and it is updated monthly with the

new microarrays from GEO. Thus, researchers can see whether the marker genes of

their microarray are marker genes in other microarrays in the database, expanding

the analysis of their microarray to the rest of the public microarrays. This

solution helps not only to corroborate the conclusions regarding a researcher's

microarray but also to identify the phenotype of different subsets of individuals

under investigation, to frame the results with microarray experiments from other

species, pathologies or tissues, to search for drugs that promote the transition

between the studied phenotypes, to detect undesirable side effects of the

treatment applied, etc. Thus, the researcher can quickly add relevant information

to his/her studies from all of the previous analyses performed in other studies

as long as they have been deposited in public repositories.

AVAILABILITY: Marker-gene database tool: http://ibb.uab.es/mgdb

© The Author 2014. Published by Oxford University Press.

DOI: 10.1093/bioinformatics/btu109

PMCID: PMC4058934

PMID: 24574116 [Indexed for MEDLINE]

824. Bioinformatics. 2014 Jun 15;30(12):1681-9. doi: 10.1093/bioinformatics/btu106.

Epub 2014 Feb 25.

Incorporating post-translational modifications and unnatural amino acids into

high-throughput modeling of protein structures.

Nagata K(1), Randall A(1), Baldi P(1).

Author information:

(1)Department of Computer Science and Institute for Genomics and Bioinformatics,

University of California, Irvine, CA, USADepartment of Computer Science and

Institute for Genomics and Bioinformatics, University of California, Irvine, CA,

USA.

MOTIVATION: Accurately predicting protein side-chain conformations is an

important subproblem of the broader protein structure prediction problem. Several

methods exist for generating fairly accurate models for moderate-size proteins in

seconds or less. However, a major limitation of these methods is their inability

to model post-translational modifications (PTMs) and unnatural amino acids. In

natural living systems, the chemical groups added following translation are often

critical for the function of the protein. In engineered systems, unnatural amino

acids are incorporated into proteins to explore structure-function relationships

and create novel proteins.

RESULTS: We present a new version of SIDEpro to predict the side chains of

proteins containing non-standard amino acids, including 15 of the most frequently

observed PTMs in the Protein Data Bank and all types of phosphorylation. SIDEpro

uses energy functions that are parameterized by neural networks trained from

available data. For PTMs, the [Formula: see text] and [Formula: see text]

accuracies are comparable with those obtained for the precursor amino acid, and

so are the RMSD values for the atoms shared with the precursor amino acid. In

addition, SIDEpro can accommodate any PTM or unnatural amino acid, thus providing

a flexible prediction system for high-throughput modeling of proteins beyond the

standard amino acids.

AVAILABILITY AND IMPLEMENTATION: SIDEpro programs and Web server, rotamer

libraries and data are available through the SCRATCH suite of protein structure

predictors at http://scratch.proteomics.ics.uci.edu/

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu106

PMCID: PMC4058938

PMID: 24574112 [Indexed for MEDLINE]

825. Bioinformatics. 2014 Jun 15;30(12):1769-70. doi: 10.1093/bioinformatics/btu096.

Epub 2014 Feb 14.

SKINK: a web server for string kernel based kink prediction in α-helices.

Seifert T(1), Lund A(1), Kneissl B(1), Mueller SC(1), Tautermann CS(1),

Hildebrandt A(1).

Author information:

(1)Department of Software Engineering and Bioinformatics, Johannes Gutenberg

University of Mainz, 55128 Mainz, Institute for Computer Science, Johann Wolfgang

Goethe-University Frankfurt am Main, 60054 Frankfurt am Main, Data Science,

Pharma Research and Early Development Informatics (pREDi), Roche Diagnostics

GmbH, 82377 Penzberg, Department of Human Genetics, Saarland University Faculty

of Medicine, 66421 Homburg and Lead Identification and Optimization Support,

Boehringer-Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany.

MOTIVATION: The reasons for distortions from optimal α-helical geometry are

widely unknown, but their influences on structural changes of proteins are

significant. Hence, their prediction is a crucial problem in structural

bioinformatics. Here, we present a new web server, called SKINK, for string

kernel based kink prediction. Extending our previous study, we also annotate the

most probable kink position in a given α-helix sequence.

AVAILABILITY AND IMPLEMENTATION: The SKINK web server is freely accessible at

http://biows-inf.zdv.uni-mainz.de/skink. Moreover, SKINK is a module of the BALL

software, also freely available at www.ballview.org.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu096

PMID: 24532729 [Indexed for MEDLINE]

826. Bioinformatics. 2014 Jun 15;30(12):1789-90. doi: 10.1093/bioinformatics/btu092.

Epub 2014 Feb 14.

GUILDify: a web server for phenotypic characterization of genes through

biological data integration and network-based prioritization algorithms.

Guney E(1), Garcia-Garcia J(1), Oliva B(1).

Author information:

(1)Departament de Ciencies Experimentals i de la Salut, Structural Bioinformatics

Group (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona, 08003 Catalonia, Spain.

SUMMARY: Determining genetic factors underlying various phenotypes is hindered by

the involvement of multiple genes acting cooperatively. Over the past years,

disease-gene prioritization has been central to identify genes implicated in

human disorders. Special attention has been paid on using physical interactions

between the proteins encoded by the genes to link them with diseases. Such

methods exploit the guilt-by-association principle in the protein interaction

network to uncover novel disease-gene associations. These methods rely on the

proximity of a gene in the network to the genes associated with a phenotype and

require a set of initial associations. Here, we present GUILDify, an easy-to-use

web server for the phenotypic characterization of genes. GUILDify offers a

prioritization approach based on the protein-protein interaction network where

the initial phenotype-gene associations are retrieved via free text search on

biological databases. GUILDify web server does not restrict the prioritization to

any predefined phenotype, supports multiple species and accepts user-specified

genes. It also prioritizes drugs based on the ranking of their targets,

unleashing opportunities for repurposing drugs for novel therapies.

AVAILABILITY AND IMPLEMENTATION: Available online at

http://sbi.imim.es/GUILDify.php

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu092

PMID: 24532728 [Indexed for MEDLINE]

827. Bioinformatics. 2014 Jun 15;30(12):1771-3. doi: 10.1093/bioinformatics/btu097.

Epub 2014 Feb 14.

ZDOCK server: interactive docking prediction of protein-protein complexes and

symmetric multimers.

Pierce BG(1), Wiehe K(2), Hwang H(1), Kim BH(2), Vreven T(2), Weng Z(1).

Author information:

(1)Program in Bioinformatics and Integrative Biology, University of Massachusetts

Medical School, 364 Plantation Street, Worcester, MA 01605 and Bioinformatics

Program, Boston University, 44 Cummington Mall, Boston, MA 02215 USAProgram in

Bioinformatics and Integrative Biology, University of Massachusetts Medical

School, 364 Plantation Street, Worcester, MA 01605 and Bioinformatics Program,

Boston University, 44 Cummington Mall, Boston, MA 02215 USA. (2)Program in

Bioinformatics and Integrative Biology, University of Massachusetts Medical

School, 364 Plantation Street, Worcester, MA 01605 and Bioinformatics Program,

Boston University, 44 Cummington Mall, Boston, MA 02215 USA.

SUMMARY: Protein-protein interactions are essential to cellular and immune

function, and in many cases, because of the absence of an experimentally

determined structure of the complex, these interactions must be modeled to obtain

an understanding of their molecular basis. We present a user-friendly protein

docking server, based on the rigid-body docking programs ZDOCK and M-ZDOCK, to

predict structures of protein-protein complexes and symmetric multimers. With a

goal of providing an accessible and intuitive interface, we provide options for

users to guide the scoring and the selection of output models, in addition to

dynamic visualization of input structures and output docking models. This server

enables the research community to easily and quickly produce structural models of

protein-protein complexes and symmetric multimers for their own analysis.

AVAILABILITY: The ZDOCK server is freely available to all academic and non-profit

users at: http://zdock.umassmed.edu. No registration is required.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu097

PMCID: PMC4058926

PMID: 24532726 [Indexed for MEDLINE]

828. BMC Bioinformatics. 2014 Jun 9;15:176. doi: 10.1186/1471-2105-15-176.

QMachine: commodity supercomputing in web browsers.

Wilkinson SR(1), Almeida JS.

Author information:

(1)Division of Informatics, Department of Pathology, University of Alabama at

Birmingham, Birmingham, USA. wilkinson@uab.edu.

BACKGROUND: Ongoing advancements in cloud computing provide novel opportunities

in scientific computing, especially for distributed workflows. Modern web

browsers can now be used as high-performance workstations for querying,

processing, and visualizing genomics' "Big Data" from sources like The Cancer

Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) without

local software installation or configuration. The design of QMachine (QM) was

driven by the opportunity to use this pervasive computing model in the context of

the Web of Linked Data in Biomedicine.

RESULTS: QM is an open-sourced, publicly available web service that acts as a

messaging system for posting tasks and retrieving results over HTTP. The

illustrative application described here distributes the analyses of 20

Streptococcus pneumoniae genomes for shared suffixes. Because all analytical and

data retrieval tasks are executed by volunteer machines, few server resources are

required. Any modern web browser can submit those tasks and/or volunteer to

execute them without installing any extra plugins or programs. A client library

provides high-level distribution templates including MapReduce. This stark

departure from the current reliance on expensive server hardware running

"download and install" software has already gathered substantial community

interest, as QM received more than 2.2 million API calls from 87 countries in 12

months.

CONCLUSIONS: QM was found adequate to deliver the sort of scalable bioinformatics

solutions that computation- and data-intensive workflows require. Paradoxically,

the sandboxed execution of code by web browsers was also found to enable them, as

compute nodes, to address critical privacy concerns that characterize biomedical

environments.

DOI: 10.1186/1471-2105-15-176

PMCID: PMC4063228

PMID: 24913605 [Indexed for MEDLINE]

829. PLoS One. 2014 Jun 9;9(6):e99368. doi: 10.1371/journal.pone.0099368. eCollection

2014.

CELLO2GO: a web server for protein subCELlular LOcalization prediction with

functional gene ontology annotation.

Yu CS(1), Cheng CW(2), Su WC(2), Chang KC(2), Huang SW(3), Hwang JK(4), Lu CH(5).

Author information:

(1)Department of Information Engineering and Computer Science, Feng Chia

University, Taichung, Taiwan; Master's Program in Biomedical Informatics and

Biomedical Engineering, Feng Chia University, Taichung, Taiwan. (2)Department of

Information Engineering and Computer Science, Feng Chia University, Taichung,

Taiwan. (3)Department of Medical Informatics, Tzu Chi University, Hualien,

Taiwan. (4)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu, Taiwan; Center of Bioinformatics Research, National Chiao

Tung University, Hsinchu, Taiwan. (5)Graduate Institute of Basic Medical Science,

China Medical University, Taichung, Taiwan.

CELLO2GO (http://cello.life.nctu.edu.tw/cello2go/) is a publicly available,

web-based system for screening various properties of a targeted protein and its

subcellular localization. Herein, we describe how this platform is used to obtain

a brief or detailed gene ontology (GO)-type categories, including subcellular

localization(s), for the queried proteins by combining the CELLO

localization-predicting and BLAST homology-searching approaches. Given a query

protein sequence, CELLO2GO uses BLAST to search for homologous sequences that are

GO annotated in an in-house database derived from the UniProt KnowledgeBase

database. At the same time, CELLO attempts predict at least one subcellular

localization on the basis of the species in which the protein is found. When

homologs for the query sequence have been identified, the number of terms found

for each of their GO categories, i.e., cellular compartment, molecular function,

and biological process, are summed and presented as pie charts representing

possible functional annotations for the queried protein. Although the

experimental subcellular localization of a protein may not be known, and thus not

annotated, CELLO can confidentially suggest a subcellular localization. CELLO2GO

should be a useful tool for research involving complex subcellular systems

because it combines CELLO and BLAST into one platform and its output is easily

manipulated such that the user-specific questions may be readily addressed.

DOI: 10.1371/journal.pone.0099368

PMCID: PMC4049835

PMID: 24911789 [Indexed for MEDLINE]

830. PLoS One. 2014 Jun 4;9(6):e98345. doi: 10.1371/journal.pone.0098345. eCollection

2014.

Protein sub-nuclear localization prediction using SVM and Pfam domain

information.

Kumar R(1), Jain S(1), Kumari B(1), Kumar M(1).

Author information:

(1)Department of Biophysics, University of Delhi South Campus, New Delhi, India.

The nucleus is the largest and the highly organized organelle of eukaryotic

cells. Within nucleus exist a number of pseudo-compartments, which are not

separated by any membrane, yet each of them contains only a specific set of

proteins. Understanding protein sub-nuclear localization can hence be an

important step towards understanding biological functions of the nucleus. Here we

have described a method, SubNucPred developed by us for predicting the

sub-nuclear localization of proteins. This method predicts protein localization

for 10 different sub-nuclear locations sequentially by combining presence or

absence of unique Pfam domain and amino acid composition based SVM model. The

prediction accuracy during leave-one-out cross-validation for centromeric

proteins was 85.05%, for chromosomal proteins 76.85%, for nuclear speckle

proteins 81.27%, for nucleolar proteins 81.79%, for nuclear envelope proteins

79.37%, for nuclear matrix proteins 77.78%, for nucleoplasm proteins 76.98%, for

nuclear pore complex proteins 88.89%, for PML body proteins 75.40% and for

telomeric proteins it was 83.33%. Comparison with other reported methods showed

that SubNucPred performs better than existing methods. A web-server for

predicting protein sub-nuclear localization named SubNucPred has been established

at http://14.139.227.92/mkumar/subnucpred/. Standalone version of SubNucPred can

also be downloaded from the web-server.

DOI: 10.1371/journal.pone.0098345

PMCID: PMC4045734

PMID: 24897370 [Indexed for MEDLINE]

831. Amino Acids. 2014 Jun;46(6):1459-69. doi: 10.1007/s00726-014-1711-5. Epub 2014

Mar 13.

PhosphoSVM: prediction of phosphorylation sites by integrating various protein

sequence attributes with a support vector machine.

Dou Y(1), Yao B, Zhang C.

Author information:

(1)Center for Plant Science and Innovation, School of Biological Sciences,

University of Nebraska, Lincoln, NE, 68588, USA.

Phosphorylation is one of the most essential post-translational modifications in

eukaryotes. Studies on kinases and their substrates are important for

understanding cellular signaling networks. Because of the cost in time and labor

associated with large-scale wet-bench experiments, computational prediction of

phosphorylation sites becomes important and many computational tools have been

developed in the recent decades. The prediction tools can be grouped into two

categories: kinase-specific and non-kinase-specific tools. With more kinases

being discovered by the new sequencing technologies, accurate non-kinase-specific

prediction tools are highly desirable for whole-genome annotation in a wider

variety of species. In this manuscript, a support vector machine is used to

combine eight different sequence level scoring functions to predict

phosphorylation sites. The attributes used by this work, including Shannon

entropy, relative entropy, predicted protein secondary structure, predicted

protein disorder, solvent accessible area, overlapping properties, averaged

cumulative hydrophobicity, and k-nearest neighbor, were able to obtain better

results than the previously used attributes by other similar methods. This method

achieved AUC values of 0.8405/0.8183/0.7383 for serine (S), threonine (T), and

tyrosine (Y) phosphorylation sites, respectively, in animals with a tenfold

cross-validation. The model trained by the animal phosphorylation sites was also

applied to a plant phosphorylation site dataset as an independent test. The AUC

values for the independent test dataset were 0.7761/0.6652/0.5958 for S/T/Y

phosphorylation sites, which compared favorably with those of several existing

methods. A web server based on our method was constructed for public use. The

server, trained model, and all datasets used in the current study are available

at http://sysbio.unl.edu/PhosphoSVM .

DOI: 10.1007/s00726-014-1711-5

PMID: 24623121 [Indexed for MEDLINE]

832. Bioinformatics. 2014 Jun 1;30(11):1522-9. doi: 10.1093/bioinformatics/btu083.

Epub 2014 Feb 6.

iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in

genomes with pseudo k-tuple nucleotide composition.

Guo SH(1), Deng EZ(1), Xu LQ(1), Ding H(1), Lin H(2), Chen W(2), Chou KC(2).

Author information:

(1)Key Laboratory for Neuro-Information of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu 610054, China, Gordon Life Science

Institute, Belmont, Massachusetts, USA, Department of Physics, School of

Sciences, Center for Genomics and Computational Biology, Hebei United University,

Tangshan 063000, China and Center of Excellence in Genomic Medicine Research

(CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia. (2)Key Laboratory for

Neuro-Information of Ministry of Education, Center of Bioinformatics, School of

Life Science and Technology, University of Electronic Science and Technology of

China, Chengdu 610054, China, Gordon Life Science Institute, Belmont,

Massachusetts, USA, Department of Physics, School of Sciences, Center for

Genomics and Computational Biology, Hebei United University, Tangshan 063000,

China and Center of Excellence in Genomic Medicine Research (CEGMR), King

Abdulaziz University, Jeddah, Saudi ArabiaKey Laboratory for Neuro-Information of

Ministry of Education, Center of Bioinformatics, School of Life Science and

Technology, University of Electronic Science and Technology of China, Chengdu

610054, China, Gordon Life Science Institute, Belmont, Massachusetts, USA,

Department of Physics, School of Sciences, Center for Genomics and Computational

Biology, Hebei United University, Tangshan 063000, China and Center of Excellence

in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi

Arabia.

MOTIVATION: Nucleosome positioning participates in many cellular activities and

plays significant roles in regulating cellular processes. With the avalanche of

genome sequences generated in the post-genomic age, it is highly desired to

develop automated methods for rapidly and effectively identifying nucleosome

positioning. Although some computational methods were proposed, most of them were

species specific and neglected the intrinsic local structural properties that

might play important roles in determining the nucleosome positioning on a DNA

sequence.

RESULTS: Here a predictor called 'iNuc-PseKNC' was developed for predicting

nucleosome positioning in Homo sapiens, Caenorhabditis elegans and Drosophila

melanogaster genomes, respectively. In the new predictor, the samples of DNA

sequences were formulated by a novel feature-vector called 'pseudo k-tuple

nucleotide composition', into which six DNA local structural properties were

incorporated. It was observed by the rigorous cross-validation tests on the three

stringent benchmark datasets that the overall success rates achieved by

iNuc-PseKNC in predicting the nucleosome positioning of the aforementioned three

genomes were 86.27%, 86.90% and 79.97%, respectively. Meanwhile, the results

obtained by iNuc-PseKNC on various benchmark datasets used by the previous

investigators for different genomes also indicated that the current predictor

remarkably outperformed its counterparts.

AVAILABILITY: A user-friendly web-server, iNuc-PseKNC is freely accessible at

http://lin.uestc.edu.cn/server/iNuc-PseKNC.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please email: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu083

PMID: 24504871 [Indexed for MEDLINE]

833. Comput Biol Chem. 2014 Jun;50:11-8. doi: 10.1016/j.compbiolchem.2014.01.009. Epub

2014 Jan 24.

Identification and characterization of lysine-methylated sites on histones and

non-histone proteins.

Lee TY(1), Chang CW(2), Lu CT(1), Cheng TH(1), Chang TH(3).

Author information:

(1)Department of Computer Science and Engineering, Yuan Ze University, Chung-Li,

Taiwan. (2)Department of Information Management, Hsing Wu University, New Taipei

City, Taiwan. (3)Graduate Institute of Biomedical Informatics, College of Medical

Science and Technology, Taipei Medical University, Taipei, Taiwan. Electronic

address: kevinchang@tmu.edu.tw.

Protein methylation is a kind of post-translational modification (PTM), and

typically takes place on lysine and arginine amino acid residues. Protein

methylation is involved in many important biological processes, and most recent

studies focused on lysine methylation of histones due to its critical roles in

regulating transcriptional repression and activation. Histones possess highly

conserved sequences and are homologous in most species. However, there is much

less sequence conservation among non-histone proteins. Therefore, mechanisms for

identifying lysine-methylated sites may greatly differ between histones and

non-histone proteins. Nevertheless, this point of view was not considered in

previous studies. Here we constructed two support vector machine (SVM) models by

using lysine-methylated data from histones and non-histone proteins for

predictions of lysine-methylated sites. Numerous features, such as the amino acid

composition (AAC) and accessible surface area (ASA), were used in the SVM models,

and the predictive performance was evaluated using five-fold cross-validations.

For histones, the predictive sensitivity was 85.62% and specificity was 80.32%.

For non-histone proteins, the predictive sensitivity was 69.1% and specificity

was 88.72%. Results showed that our model significantly improved the predictive

accuracy of histones compared to previous approaches. In addition, features of

the flanking region of lysine-methylated sites on histones and non-histone

proteins were also characterized and are discussed. A gene ontology functional

analysis of lysine-methylated proteins and correlations of lysine-methylated

sites with other PTMs in histones were also analyzed in detail. Finally, a web

server, MethyK, was constructed to identify lysine-methylated sites. MethK now is

available at http://csb.cse.yzu.edu.tw/MethK/.

Copyright © 2014 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiolchem.2014.01.009

PMID: 24560580 [Indexed for MEDLINE]

834. J Mol Model. 2014 Jun;20(6):2278. doi: 10.1007/s00894-014-2278-5. Epub 2014 May

31.

AllerTOP v.2--a server for in silico prediction of allergens.

Dimitrov I(1), Bangov I, Flower DR, Doytchinova I.

Author information:

(1)Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st., 1000, Sofia,

Bulgaria.

Allergy is an overreaction by the immune system to a previously encountered,

ordinarily harmless substance--typically proteins--resulting in skin rash,

swelling of mucous membranes, sneezing or wheezing, or other abnormal conditions.

The use of modified proteins is increasingly widespread: their presence in food,

commercial products, such as washing powder, and medical therapeutics and

diagnostics, makes predicting and identifying potential allergens a crucial

societal issue. The prediction of allergens has been explored widely using

bioinformatics, with many tools being developed in the last decade; many of these

are freely available online. Here, we report a set of novel models for allergen

prediction utilizing amino acid E-descriptors, auto- and cross-covariance

transformation, and several machine learning methods for classification,

including logistic regression (LR), decision tree (DT), naïve Bayes (NB), random

forest (RF), multilayer perceptron (MLP) and k nearest neighbours (kNN). The best

performing method was kNN with 85.3% accuracy at 5-fold cross-validation. The

resulting model has been implemented in a revised version of the AllerTOP server

(http://www.ddg-pharmfac.net/AllerTOP).

DOI: 10.1007/s00894-014-2278-5

PMID: 24878803 [Indexed for MEDLINE]

835. Methods. 2014 Jun 1;67(3):386-93. doi: 10.1016/j.ymeth.2014.01.008. Epub 2014 Jan

15.

Literature mining of protein phosphorylation using dependency parse trees.

Wang M(1), Xia H(2), Sun D(3), Chen Z(4), Wang M(5), Li A(6).

Author information:

(1)School of Information Science and Technology, University of Science and

Technology of China, Hefei AH230027, China. Electronic address:

mangwang@mail.ustc.edu.cn. (2)School of Information Science and Technology,

University of Science and Technology of China, Hefei AH230027, China. Electronic

address: sommer@mail.ustc.edu.cn. (3)School of Information Science and

Technology, University of Science and Technology of China, Hefei AH230027, China.

Electronic address: sddchina@mail.ustc.edu.cn. (4)School of Life Sciences,

University of Science and Technology of China, Hefei AH230027, China. Electronic

address: czx01@mail.ustc.edu.cn. (5)School of Information Science and Technology,

University of Science and Technology of China, Hefei AH230027, China; Centers for

Biomedical Engineering, University of Science and Technology of China, Hefei

AH230027, China. Electronic address: mhwang@ustc.edu.cn. (6)School of Information

Science and Technology, University of Science and Technology of China, Hefei

AH230027, China; Centers for Biomedical Engineering, University of Science and

Technology of China, Hefei AH230027, China. Electronic address: aoli@ustc.edu.cn.

As one of the most common post-translational modifications (PTMs), protein

phosphorylation plays an important role in various biological processes, such as

signaling transduction, cellular metabolism, differentiation, growth, regulation

and apoptosis. Protein phosphorylation is of great value not only in illustrating

the underlying molecular mechanisms but also in treatment of diseases and design

of new drugs. Recently, there is an increasing interest in automatically

extracting phosphorylation information from biomedical literatures. However, it

still remains a challenging task due to the tremendous volume of literature and

circuitous modes of expression for protein phosphorylation. To address this

issue, we propose a novel text-mining method for efficiently retrieving and

extracting protein phosphorylation information from literature. By employing

natural language processing (NLP) technologies, this method transforms each

sentence into dependency parse trees that can precisely reflect the intrinsic

relationship of phosphorylation-related key words, from which detailed

information of substrates, kinases and phosphorylation sites is extracted based

on syntactic patterns. Compared with other existing approaches, the proposed

method demonstrates significantly improved performance, suggesting it is a

powerful bioinformatics approach to retrieving phosphorylation information from a

large amount of literature. A web server for the proposed method is freely

available at http://bioinformatics.ustc.edu.cn/pptm/.

Copyright © 2014 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ymeth.2014.01.008

PMID: 24440484 [Indexed for MEDLINE]

836. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi. 2014 Jun;31(3):563-6.

[WEB-based medical data mining integration].

[Article in Chinese]

Yao G, Zhang X, Wang H.

An integration of medical data management system based on WEB and data mining

tool is reportedly in this paper. In the application process of this system,

web-based medical data mining user sends requests to the server by using client

browser with http protocol, the commands are then received by the server and the

server calls the data mining tools remote object for data processing, and the

results are sent back to the customer browser through the http protocol and

presented to the user. In order to prove the feasibility of the proposed

solution, the test is done under the NET platform by using SAS and SPSS, and the

detail steps are given. By the practical test, it was proved that the web-based

data mining tool integration solutions proposed in this paper would have its

broad prospects for development, which would open up a new route to the

development of medical data mining.

PMID: 25219235 [Indexed for MEDLINE]

837. Trends Parasitol. 2014 Jun;30(6):309-16. doi: 10.1016/j.pt.2014.04.006. Epub 2014

Apr 30.

MalarImDB: an open-access literature-based malaria immunology database.

Deroost K(1), Opdenakker G(1), Van den Steen PE(2).

Author information:

(1)Laboratory of Immunobiology, Rega Institute for Medical Research, KU Leuven -

University of Leuven, Minderbroedersstraat 10, 3000 Leuven, Belgium.

(2)Laboratory of Immunobiology, Rega Institute for Medical Research, KU Leuven -

University of Leuven, Minderbroedersstraat 10, 3000 Leuven, Belgium. Electronic

address: Philippe.vandensteen@rega.kuleuven.be.

The Malaria Immunology Database (MalarImDB, www.malarimdb.org) is a novel

literature-based database of host mediators in blood-stage malaria. We designed

this open-access online tool because intensive malaria research has resulted in a

dazzling complexity of host mediators with pathogenic or protective functions.

MalarImDB allows comparisons between expression levels in humans, expression

levels in murine models, and functional data from experimental treatments in

mice. The database is equipped with multiple search engines to retrieve

information from many published studies. The search output is visualized

schematically in tables, thereby revealing similarities and disparities. Thus,

the primary aim of this database is to present a clear overview of the currently

available data about malaria and to simplify literature searches.

Copyright © 2014 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.pt.2014.04.006

PMID: 24794510 [Indexed for MEDLINE]

838. BMC Genomics. 2014 May 29;15:411. doi: 10.1186/1471-2164-15-411.

Predicting the fungal CUG codon translation with Bagheera.

Mühlhausen S, Kollmar M(1).

Author information:

(1)Group Systems Biology of Motor Proteins, Department of NMR-based Structural

Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany.

mako@nmr.mpibpc.mpg.de.

BACKGROUND: Many eukaryotes have been shown to use alternative schemes to the

universal genetic code. While most Saccharomycetes, including Saccharomyces

cerevisiae, use the standard genetic code translating the CUG codon as leucine,

some yeasts, including many but not all of the "Candida", translate the same

codon as serine. It has been proposed that the change in codon identity was

accomplished by an almost complete loss of the original CUG codons, making the

CUG positions within the extant species highly discriminative for the one or

other translation scheme.

RESULTS: In order to improve the prediction of genes in yeast species by

providing the correct CUG decoding scheme we implemented a web server, called

Bagheera, that allows determining the most probable CUG codon translation for a

given transcriptome or genome assembly based on extensive reference data. As

reference data we use 2071 manually assembled and annotated sequences from 38

cytoskeletal and motor proteins belonging to 79 yeast species. The web service

includes a pipeline, which starts with predicting and aligning homologous genes

to the reference data. CUG codon positions within the predicted genes are

analysed with respect to amino acid similarity and CUG codon conservation in

related species. In addition, the tRNACAG gene is predicted in genomic data and

compared to known leu-tRNACAG and ser-tRNACAG genes. Bagheera can also be used to

evaluate any mRNA and protein sequence data with the codon usage of the

respective species. The usage of the system has been demonstrated by analysing

six genomes not included in the reference data.

CONCLUSIONS: Gene prediction and consecutive comparison with reference data from

other Saccharomycetes are sufficient to predict the most probable decoding scheme

for CUG codons. This approach has been implemented into Bagheera

(http://www.motorprotein.de/bagheera).

DOI: 10.1186/1471-2164-15-411

PMCID: PMC4050208

PMID: 24885275 [Indexed for MEDLINE]

839. J Cheminform. 2014 May 23;6:28. doi: 10.1186/1758-2946-6-28. eCollection 2014.

iDrug: a web-accessible and interactive drug discovery and design platform.

Wang X(1), Chen H(2), Yang F(2), Gong J(2), Li S(1), Pei J(3), Liu X(1), Jiang

H(1), Lai L(3), Li H(4).

Author information:

(1)Shanghai Key Laboratory of New Drug Design, State Key Laboratory of Bioreactor

Engineering, School of Pharmacy, East China University of Science and Technology,

Shanghai 200237, China. (2)School of Information Science and Engineering, East

China University of Science and Technology, Shanghai 200237, China. (3)BNLMS,

Center for Quantitative Biology, State Key Laboratory for Structural Chemistry of

Unstable and Stable Species, College of Chemistry and Molecular Engineering,

Peking University, Beijing 100871, China. (4)Shanghai Key Laboratory of New Drug

Design, State Key Laboratory of Bioreactor Engineering, School of Pharmacy, East

China University of Science and Technology, Shanghai 200237, China ; School of

Information Science and Engineering, East China University of Science and

Technology, Shanghai 200237, China.

BACKGROUND: The progress in computer-aided drug design (CADD) approaches over the

past decades accelerated the early-stage pharmaceutical research. Many powerful

standalone tools for CADD have been developed in academia. As programs are

developed by various research groups, a consistent user-friendly online graphical

working environment, combining computational techniques such as pharmacophore

mapping, similarity calculation, scoring, and target identification is needed.

RESULTS: We presented a versatile, user-friendly, and efficient online tool for

computer-aided drug design based on pharmacophore and 3D molecular similarity

searching. The web interface enables binding sites detection, virtual screening

hits identification, and drug targets prediction in an interactive manner through

a seamless interface to all adapted packages (e.g., Cavity, PocketV.2,

PharmMapper, SHAFTS). Several commercially available compound databases for hit

identification and a well-annotated pharmacophore database for drug targets

prediction were integrated in iDrug as well. The web interface provides tools for

real-time molecular building/editing, converting, displaying, and analyzing. All

the customized configurations of the functional modules can be accessed through

featured session files provided, which can be saved to the local disk and

uploaded to resume or update the history work.

CONCLUSIONS: iDrug is easy to use, and provides a novel, fast and reliable tool

for conducting drug design experiments. By using iDrug, various molecular design

processing tasks can be submitted and visualized simply in one browser without

installing locally any standalone modeling softwares. iDrug is accessible free of

charge at http://lilab.ecust.edu.cn/idrug.

DOI: 10.1186/1758-2946-6-28

PMCID: PMC4046018

PMID: 24955134

840. Front Neuroinform. 2014 May 21;8:52. doi: 10.3389/fninf.2014.00052. eCollection

2014.

A simple tool for neuroimaging data sharing.

Haselgrove C(1), Poline JB(1), Kennedy DN(1).

Author information:

(1)University of Massachusetts Medical School Worcester, MA, USA.

Data sharing is becoming increasingly common, but despite encouragement and

facilitation by funding agencies, journals, and some research efforts, most

neuroimaging data acquired today is still not shared due to political, financial,

social, and technical barriers to sharing data that remain. In particular,

technical solutions are few for researchers that are not a part of larger efforts

with dedicated sharing infrastructures, and social barriers such as the time

commitment required to share can keep data from becoming publicly available. We

present a system for sharing neuroimaging data, designed to be simple to use and

to provide benefit to the data provider. The system consists of a server at the

International Neuroinformatics Coordinating Facility (INCF) and user tools for

uploading data to the server. The primary design principle for the user tools is

ease of use: the user identifies a directory containing Digital Imaging and

Communications in Medicine (DICOM) data, provides their INCF Portal

authentication, and provides identifiers for the subject and imaging session. The

user tool anonymizes the data and sends it to the server. The server then runs

quality control routines on the data, and the data and the quality control

reports are made public. The user retains control of the data and may change the

sharing policy as they need. The result is that in a few minutes of the user's

time, DICOM data can be anonymized and made publicly available, and an initial

quality control assessment can be performed on the data. The system is currently

functional, and user tools and access to the public image database are available

at http://xnat.incf.org/.

DOI: 10.3389/fninf.2014.00052

PMCID: PMC4033259

PMID: 24904398

841. PLoS One. 2014 May 14;9(5):e97158. doi: 10.1371/journal.pone.0097158. eCollection

2014.

Propensity scores for prediction and characterization of bioluminescent proteins

from sequences.

Huang HL(1).

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu, Taiwan; Department of Biological Science and Technology,

National Chiao Tung University, Hsinchu, Taiwan.

Bioluminescent proteins (BLPs) are a class of proteins with various mechanisms of

light emission such as bioluminescence and fluorescence from luminous organisms.

While valuable for commercial and medical applications, identification of BLPs,

including luciferases and fluorescent proteins (FPs), is rather challenging,

owing to their high variety of protein sequences. Moreover, characterization of

BLPs facilitates mutagenesis analysis to enhance bioluminescence and

fluorescence. Therefore, this study proposes a novel methodological approach to

estimating the propensity scores of 400 dipeptides and 20 amino acids in order to

design two prediction methods and characterize BLPs based on a scoring card

method (SCM). The SCMBLP method for predicting BLPs achieves an accuracy of

90.83% for 10-fold cross-validation higher than existing support vector machine

based methods and a test accuracy of 82.85%. A dataset consisting of 269

luciferases and 216 FPs is also established to design the SCMLFP prediction

method, which achieves training and test accuracies of 97.10% and 96.28%,

respectively. Additionally, four informative physicochemical properties of 20

amino acids are identified using the estimated propensity scores to characterize

BLPs as follows: 1) high transfer free energy from inside to the protein surface,

2) high occurrence frequency of residues in the transmembrane regions of the

protein, 3) large hydrophobicity scale from the native protein structure, and 4)

high correlation coefficient (R = 0.921) between the amino acid compositions of

BLPs and integral membrane proteins. Further analyzing BLPs reveals that

luciferases have a larger value of R (0.937) than FPs (0.635), suggesting that

luciferases tend to locate near the cell membrane location rather than FPs for

convenient receipt of extracellular ions. Importantly, the propensity scores of

dipeptides and amino acids and the identified properties facilitate efforts to

predict, characterize, and apply BLPs, including luciferases, photoproteins, and

FPs. The web server is available at

http://iclab.life.nctu.edu.tw/SCMBLP/index.html.

DOI: 10.1371/journal.pone.0097158

PMCID: PMC4020813

PMID: 24828431 [Indexed for MEDLINE]

842. PLoS One. 2014 May 14;9(5):e97446. doi: 10.1371/journal.pone.0097446. eCollection

2014.

Supervised learning classification models for prediction of plant virus encoded

RNA silencing suppressors.

Jagga Z(1), Gupta D(1).

Author information:

(1)Bioinformatics Facility, Structural and Computational Biology Group,

International Centre for Genetic Engineering and Biotechnology, Aruna Asaf Ali

Marg, New Delhi, India.

Viral encoded RNA silencing suppressor proteins interfere with the host RNA

silencing machinery, facilitating viral infection by evading host immunity. In

plant hosts, the viral proteins have several basic science implications and

biotechnology applications. However in silico identification of these proteins is

limited by their high sequence diversity. In this study we developed supervised

learning based classification models for plant viral RNA silencing suppressor

proteins in plant viruses. We developed four classifiers based on supervised

learning algorithms: J48, Random Forest, LibSVM and Naïve Bayes algorithms, with

enriched model learning by correlation based feature selection. Structural and

physicochemical features calculated for experimentally verified primary protein

sequences were used to train the classifiers. The training features include amino

acid composition; auto correlation coefficients; composition, transition, and

distribution of various physicochemical properties; and pseudo amino acid

composition. Performance analysis of predictive models based on 10 fold

cross-validation and independent data testing revealed that the Random Forest

based model was the best and achieved 86.11% overall accuracy and 86.22% balanced

accuracy with a remarkably high area under the Receivers Operating Characteristic

curve of 0.95 to predict viral RNA silencing suppressor proteins. The prediction

models for plant viral RNA silencing suppressors can potentially aid

identification of novel viral RNA silencing suppressors, which will provide

valuable insights into the mechanism of RNA silencing and could be further

explored as potential targets for designing novel antiviral therapeutics. Also,

the key subset of identified optimal features may help in determining

compositional patterns in the viral proteins which are important determinants for

RNA silencing suppressor activities. The best prediction model developed in the

study is available as a freely accessible web server pVsupPred at

http://bioinfo.icgeb.res.in/pvsup/.

DOI: 10.1371/journal.pone.0097446

PMCID: PMC4020838

PMID: 24828116 [Indexed for MEDLINE]

843. PLoS One. 2014 May 12;9(5):e95221. doi: 10.1371/journal.pone.0095221. eCollection

2014.

Predicting mTOR inhibitors with a classifier using recursive partitioning and

Naïve Bayesian approaches.

Wang L(1), Chen L(1), Liu Z(1), Zheng M(1), Gu Q(1), Xu J(1).

Author information:

(1)Research Center for Drug Discovery & Institute of Human Virology, School of

Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou, China.

BACKGROUND: Mammalian target of rapamycin (mTOR) is a central controller of cell

growth, proliferation, metabolism, and angiogenesis. Thus, there is a great deal

of interest in developing clinical drugs based on mTOR. In this paper, in silico

models based on multi-scaffolds were developed to predict mTOR inhibitors or

non-inhibitors.

METHODS: First 1,264 diverse compounds were collected and categorized as mTOR

inhibitors and non-inhibitors. Two methods, recursive partitioning (RP) and naïve

Bayesian (NB), were used to build combinatorial classification models of mTOR

inhibitors versus non-inhibitors using physicochemical descriptors, fingerprints,

and atom center fragments (ACFs).

RESULTS: A total of 253 models were constructed and the overall predictive

accuracies of the best models were more than 90% for both the training set of 964

and the external test set of 300 diverse compounds. The scaffold hopping

abilities of the best models were successfully evaluated through predicting 37

new recently published mTOR inhibitors. Compared with the best RP and Bayesian

models, the classifier based on ACFs and Bayesian shows comparable or slightly

better in performance and scaffold hopping abilities. A web server was developed

based on the ACFs and Bayesian method (http://rcdd.sysu.edu.cn/mtor/). This web

server can be used to predict whether a compound is an mTOR inhibitor or

non-inhibitor online.

CONCLUSION: In silico models were constructed to predict mTOR inhibitors using

recursive partitioning and naïve Bayesian methods, and a web server (mTOR

Predictor) was also developed based on the best model results. Compound

prediction or virtual screening can be carried out through our web server.

Moreover, the favorable and unfavorable fragments for mTOR inhibitors obtained

from Bayesian classifiers will be helpful for lead optimization or the design of

new mTOR inhibitors.

DOI: 10.1371/journal.pone.0095221

PMCID: PMC4018356

PMID: 24819222 [Indexed for MEDLINE]

844. BMC Genomics. 2014 May 7;15:344. doi: 10.1186/1471-2164-15-344.

Genome-wide identification of heat shock proteins (Hsps) and Hsp interactors in

rice: Hsp70s as a case study.

Wang Y, Lin S, Song Q, Li K, Tao H, Huang J, Chen X, Que S, He H(1).

Author information:

(1)College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou

350002, China. hehq16@gmail.com.

BACKGROUND: Heat shock proteins (Hsps) perform a fundamental role in protecting

plants against abiotic stresses. Although researchers have made great efforts on

the functional analysis of individual family members, Hsps have not been fully

characterized in rice (Oryza sativa L.) and little is known about their

interactors.

RESULTS: In this study, we combined orthology-based approach with expression

association data to screen rice Hsps for the expression patterns of which

strongly correlated with that of heat responsive probe-sets. Twenty-seven Hsp

candidates were identified, including 12 small Hsps, six Hsp70s, three Hsp60s,

three Hsp90s, and three clpB/Hsp100s. Then, using a combination of interolog and

expression profile-based methods, we inferred 430 interactors of Hsp70s in rice,

and validated the interactions by co-localization and function-based methods.

Subsequent analysis showed 13 interacting domains and 28 target motifs were

over-represented in Hsp70s interactors. Twenty-four GO terms of biological

processes and five GO terms of molecular functions were enriched in the positive

interactors, whose expression levels were positively associated with Hsp70s.

Hsp70s interaction network implied that Hsp70s were involved in macromolecular

translocation, carbohydrate metabolism, innate immunity, photosystem II repair

and regulation of kinase activities.

CONCLUSIONS: Twenty-seven Hsps in rice were identified and 430 interactors of

Hsp70s were inferred and validated, then the interacting network of Hsp70s was

induced and the function of Hsp70s was analyzed. Furthermore, two databases named

Rice Heat Shock Proteins (RiceHsps) and Rice Gene Expression Profile (RGEP), and

one online tool named Protein-Protein Interaction Predictor (PPIP), were

constructed and could be accessed at http://bioinformatics.fafu.edu.cn/.

DOI: 10.1186/1471-2164-15-344

PMCID: PMC4035072

PMID: 24884676 [Indexed for MEDLINE]

845. Int J Mol Sci. 2014 May 5;15(5):7594-610. doi: 10.3390/ijms15057594.

iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by

incorporating dipeptide position-specific propensity into pseudo amino acid

composition.

Xu Y(1), Wen X(2), Shao XJ(3), Deng NY(4), Chou KC(5).

Author information:

(1)Department of Information and Computer Science, University of Science and

Technology Beijing, Beijing 100083, China. xyan@gordonlifescience.org.

(2)Department of Information and Computer Science, University of Science and

Technology Beijing, Beijing 100083, China. wenxinfairy@gmail.com. (3)Department

of Mathematics and Information Science, Binzhou University, Binzhou 256603,

China. shaoxiaojian@gmail.com. (4)College of Science, China Agricultural

University, Beijing 100083, China. dengnaiyang@cau.edu.cn. (5)Center of

Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University,

Jeddah 21589, Saudi Arabia. kcchou@gordonlifescience.org.

Post-translational modifications (PTMs) play crucial roles in various cell

functions and biological processes. Protein hydroxylation is one type of PTM that

usually occurs at the sites of proline and lysine. Given an uncharacterized

protein sequence, which site of its Pro (or Lys) can be hydroxylated and which

site cannot? This is a challenging problem, not only for in-depth understanding

of the hydroxylation mechanism, but also for drug development, because protein

hydroxylation is closely relevant to major diseases, such as stomach and lung

cancers. With the avalanche of protein sequences generated in the post-genomic

age, it is highly desired to develop computational methods to address this

problem. In view of this, a new predictor called "iHyd-PseAAC" (identify

hydroxylation by pseudo amino acid composition) was proposed by incorporating the

dipeptide position-specific propensity into the general form of pseudo amino acid

composition. It was demonstrated by rigorous cross-validation tests on stringent

benchmark datasets that the new predictor is quite promising and may become a

useful high throughput tool in this area. A user-friendly web-server for

iHyd-PseAAC is accessible at http://app.aporc.org/iHyd-PseAAC/. Furthermore, for

the convenience of the majority of experimental scientists, a step-by-step guide

on how to use the web-server is given. Users can easily obtain their desired

results by following these steps without the need of understanding the

complicated mathematical equations presented in this paper just for its

integrity.

DOI: 10.3390/ijms15057594

PMCID: PMC4057693

PMID: 24857907 [Indexed for MEDLINE]

846. PLoS One. 2014 May 2;9(5):e96694. doi: 10.1371/journal.pone.0096694. eCollection

2014.

Predicting DNA-binding proteins and binding residues by complex structure

prediction and application to human proteome.

Zhao H(1), Wang J(2), Zhou Y(3), Yang Y(4).

Author information:

(1)School of Informatics, Indiana University Purdue University, Indianapolis,

Indiana, United States of America; Center for Computational Biology and

Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana,

United States of America; QIMR Berghofer Medical Research Institute, Brisbane,

Queensland, Australia. (2)Center for Computational Biology and Bioinformatics,

Indiana University School of Medicine, Indianapolis, Indiana, United States of

America; Shandong Provincial Key Laboratory of Functional Macromolecular

Biophysics, Dezhou University, Dezhou, Shandong, China. (3)School of Informatics,

Indiana University Purdue University, Indianapolis, Indiana, United States of

America; Center for Computational Biology and Bioinformatics, Indiana University

School of Medicine, Indianapolis, Indiana, United States of America; Shandong

Provincial Key Laboratory of Functional Macromolecular Biophysics, Dezhou

University, Dezhou, Shandong, China; Institute for Glycomics and School of

Information and Communication Technique, Griffith University, Southport,

Queensland, Australia. (4)School of Informatics, Indiana University Purdue

University, Indianapolis, Indiana, United States of America; Center for

Computational Biology and Bioinformatics, Indiana University School of Medicine,

Indianapolis, Indiana, United States of America; Institute for Glycomics and

School of Information and Communication Technique, Griffith University,

Southport, Queensland, Australia.

As more and more protein sequences are uncovered from increasingly inexpensive

sequencing techniques, an urgent task is to find their functions. This work

presents a highly reliable computational technique for predicting DNA-binding

function at the level of protein-DNA complex structures, rather than

low-resolution two-state prediction of DNA-binding as most existing techniques

do. The method first predicts protein-DNA complex structure by utilizing the

template-based structure prediction technique HHblits, followed by binding

affinity prediction based on a knowledge-based energy function (Distance-scaled

finite ideal-gas reference state for protein-DNA interactions). A leave-one-out

cross validation of the method based on 179 DNA-binding and 3797 non-binding

protein domains achieves a Matthews correlation coefficient (MCC) of 0.77 with

high precision (94%) and high sensitivity (65%). We further found 51% sensitivity

for 82 newly determined structures of DNA-binding proteins and 56% sensitivity

for the human proteome. In addition, the method provides a reasonably accurate

prediction of DNA-binding residues in proteins based on predicted DNA-binding

complex structures. Its application to human proteome leads to more than 300

novel DNA-binding proteins; some of these predicted structures were validated by

known structures of homologous proteins in APO forms. The method [SPOT-Seq (DNA)]

is available as an on-line server at http://sparks-lab.org.

DOI: 10.1371/journal.pone.0096694

PMCID: PMC4008587

PMID: 24792350 [Indexed for MEDLINE]

847. Biochim Biophys Acta. 2014 May;1844(5):960-6. doi: 10.1016/j.bbapap.2013.11.007.

Epub 2013 Nov 20.

ProfileDB: a resource for proteomics and cross-omics biomarker discovery.

Bauer C(1), Glintschert A(1), Schuchhardt J(2).

Author information:

(1)MicroDiscovery GmbH, Marienburger Str. 1, D-10405 Berlin, Germany.

(2)MicroDiscovery GmbH, Marienburger Str. 1, D-10405 Berlin, Germany. Electronic

address: johannes.schuchhardt@microdiscovery.de.

The increasing size and complexity of high-throughput datasets pose a growing

challenge for researchers. Often very different (cross-omics) techniques with

individual data analysis pipelines are employed making a unified biomarker

discovery strategy and a direct comparison of different experiments difficult and

time consuming. Here we present the comprehensive web-based application

ProfileDB. The application is designed to integrate data from different

high-throughput 'omics' data types (Transcriptomics, Proteomics, Metabolomics)

with clinical parameters and prior knowledge on pathways and ontologies. Beyond

data storage, ProfileDB provides a set of dedicated tools for study inspection

and data visualization. The user can gain insights into a complex experiment with

just a few mouse clicks. We will demonstrate the application by presenting

typical use cases for the identification of proteomics biomarkers. All presented

analyses can be reproduced using the public ProfileDB web server. The ProfileDB

application is available by standard browser (Firefox 18+, Internet Explorer

Version 9+) technology via http://profileDB.-microdiscovery.de/ (login and

pass-word: profileDB). The installation contains several public datasets

including different cross-'omics' experiments. This article is part of a Special

Issue entitled: Biomarkers: A Proteomic Challenge.

Copyright © 2014 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.bbapap.2013.11.007

PMID: 24270047 [Indexed for MEDLINE]

848. Georgian Med News. 2014 May;(230):46-53.

Congenital diseases of the gastrointestinal tract.

Lentze M(1).

Author information:

(1)Prof. em. Pediatrics, Zentrum für Kinderheilkunde, Bonn University, Germany.

With the rapid increase in knowledge on the genetic origin of diseases within the

gastrointestinal tract the number of congenital diseases, which already manifest

during childhood have drastically increased. Due to the large application of

molecular genetics the number is steadily increasing. To make the access to these

rare diseases fast and efficient the data base of the National Library of

Medicine (Online Mendelian Inheritance of Man - OMIN) is a very helpful online

tool, with which all these disease entities can be found easily

(http://www.ncbi.nlm.nih.gov/omim). Detailed tables are given to find most of the

congenitally inherited disease, which affect the gastrointestinal tract. A

variety of congenital diarrheas with disturbances of digestion, hydrolysis,

absorption and secretion is described in detail: lactose intolerance, sucrose

intolerance, glucose-galactose malabsorption, fructose malabsorption, trehalase

and enterokinase deficiency, congenital chloride and sodium diarrhea, congenital

hypomagnesaemia, primary bile acid malabsorption, acrodermatitis enteropathica

and Menke's syndrome. Also described in detail are diseases with structural

anomalies of the intestine like microvillous inclusion disease, congenital

tufting enteropathy and IPEX syndrome. The diagnosis in the disturbances of

carbohydrate hydrolysis or absorption can be established by H2-breath tests after

appropriate sugar challenge. Treatment consists of elimination of the responsible

sugar from the diet. The diagnosis of the congenital secretory diarrheas is

established by investigation of electrolytes in blood and stool. Substitution of

high doses of the responsible mineral can improve the clinical outcome. In

acrodermatitis enteropathica low serum zinc level together with the typical skin

lesions guide to the diagnosis. High doses of oral zinc aspartate can cure the

symptoms of the disease. The diagnosis of structural congenital lesions of the

intestine can be established by histology and/or electron microscopy and

molecular identification of the responsible mutations. The treatment of these

diseases is difficult and therefore the prognosis remains poor. Immunosupressive

therapy, total parenteral nutrition and even intestinal or bone marrow

transplantation are the only choice for treatment.

PMID: 24940857 [Indexed for MEDLINE]

849. Hum Mutat. 2014 May;35(5):537-47. doi: 10.1002/humu.22520. Epub 2014 Mar 6.

Prioritizing disease-linked variants, genes, and pathways with an interactive

whole-genome analysis pipeline.

Lee IH(1), Lee K, Hsing M, Choe Y, Park JH, Kim SH, Bohn JM, Neu MB, Hwang KB,

Green RC, Kohane IS, Kong SW.

Author information:

(1)Children's Hospital Informatics Program at the Harvard-MIT Division of Health

Sciences and Technology, Department of Medicine, Boston Children's Hospital,

Boston, Massachusetts, 02115.

Whole-genome sequencing (WGS) studies are uncovering disease-associated variants

in both rare and nonrare diseases. Utilizing the next-generation sequencing for

WGS requires a series of computational methods for alignment, variant detection,

and annotation, and the accuracy and reproducibility of annotation results are

essential for clinical implementation. However, annotating WGS with up to date

genomic information is still challenging for biomedical researchers. Here, we

present one of the fastest and highly scalable annotation, filtering, and

analysis pipeline-gNOME-to prioritize phenotype-associated variants while

minimizing false-positive findings. Intuitive graphical user interface of gNOME

facilitates the selection of phenotype-associated variants, and the result

summaries are provided at variant, gene, and genome levels. Moreover, the

enrichment results of specific variants, genes, and gene sets between two groups

or compared with population scale WGS datasets that is already integrated in the

pipeline can help the interpretation. We found a small number of discordant

results between annotation software tools in part due to different reporting

strategies for the variants with complex impacts. Using two published whole-exome

datasets of uveal melanoma and bladder cancer, we demonstrated gNOME's accuracy

of variant annotation and the enrichment of loss-of-function variants in known

cancer pathways. gNOME Web server and source codes are freely available to the

academic community (http://gnome.tchlab.org).

© 2014 WILEY PERIODICALS, INC.

DOI: 10.1002/humu.22520

PMCID: PMC4130156

PMID: 24478219 [Indexed for MEDLINE]

850. Nat Protoc. 2014 May;9(5):1056-82. doi: 10.1038/nprot.2014.063. Epub 2014 Apr 10.

Characterization of ancient and modern genomes by SNP detection and phylogenomic

and metagenomic analysis using PALEOMIX.

Schubert M(1), Ermini L(1), Der Sarkissian C(1), Jónsson H(1), Ginolhac A(1),

Schaefer R(2), Martin MD(1), Fernández R(1), Kircher M(3), McCue M(4), Willerslev

E(1), Orlando L(1).

Author information:

(1)Centre for GeoGenetics, Natural History Museum of Denmark, University of

Copenhagen, Copenhagen, Denmark. (2)Biomedical Informatics and Computational

Biology Graduate Program, University of Minnesota Rochester, Rochester,

Minnesota, USA. (3)Department of Genome Sciences, University of Washington,

Seattle, Washington, USA. (4)College of Veterinary Medicine, University of

Minnesota, St. Paul, Minnesota, USA.

Next-generation sequencing technologies have revolutionized the field of

paleogenomics, allowing the reconstruction of complete ancient genomes and their

comparison with modern references. However, this requires the processing of vast

amounts of data and involves a large number of steps that use a variety of

computational tools. Here we present PALEOMIX

(http://geogenetics.ku.dk/publications/paleomix), a flexible and user-friendly

pipeline applicable to both modern and ancient genomes, which largely automates

the in silico analyses behind whole-genome resequencing. Starting with

next-generation sequencing reads, PALEOMIX carries out adapter removal, mapping

against reference genomes, PCR duplicate removal, characterization of and

compensation for postmortem damage, SNP calling and maximum-likelihood

phylogenomic inference, and it profiles the metagenomic contents of the samples.

As such, PALEOMIX allows for a series of potential applications in paleogenomics,

comparative genomics and metagenomics. Applying the PALEOMIX pipeline to the

three ancient and seven modern Phytophthora infestans genomes as described here

takes 5 d using a 16-core server.

DOI: 10.1038/nprot.2014.063

PMID: 24722405 [Indexed for MEDLINE]

851. Plant Cell Environ. 2014 May;37(5):1250-8. doi: 10.1111/pce.12231. Epub 2013 Dec

17.

Mercator: a fast and simple web server for genome scale functional annotation of

plant sequence data.

Lohse M(1), Nagel A, Herter T, May P, Schroda M, Zrenner R, Tohge T, Fernie AR,

Stitt M, Usadel B.

Author information:

(1)Max-Planck-Institute of Molecular Plant Physiology, 14476, Potsdam, Germany.

Next-generation technologies generate an overwhelming amount of gene sequence

data. Efficient annotation tools are required to make these data amenable to

functional genomics analyses. The Mercator pipeline automatically assigns

functional terms to protein or nucleotide sequences. It uses the MapMan 'BIN'

ontology, which is tailored for functional annotation of plant 'omics' data. The

classification procedure performs parallel sequence searches against reference

databases, compiles the results and computes the most likely MapMan BINs for each

query. In the current version, the pipeline relies on manually curated reference

classifications originating from the three reference organisms (Arabidopsis,

Chlamydomonas, rice), various other plant species that have a reviewed SwissProt

annotation, and more than 2000 protein domain and family profiles at InterPro,

CDD and KOG. Functional annotations predicted by Mercator achieve accuracies

above 90% when benchmarked against manual annotation. In addition to mapping

files for direct use in the visualization software MapMan, Mercator provides

graphical overview charts, detailed annotation information in a convenient web

browser interface and a MapMan-to-GO translation table to export results as GO

terms. Mercator is available free of charge via

http://mapman.gabipd.org/web/guest/app/Mercator.

© 2013 John Wiley & Sons Ltd.

DOI: 10.1111/pce.12231

PMID: 24237261 [Indexed for MEDLINE]

852. PLoS Comput Biol. 2014 May 1;10(5):e1003592. doi: 10.1371/journal.pcbi.1003592.

eCollection 2014.

Determining effects of non-synonymous SNPs on protein-protein interactions using

supervised and semi-supervised learning.

Zhao N(1), Han JG(1), Shyu CR(2), Korkin D(3).

Author information:

(1)Informatics Institute, University of Missouri, Columbia, Missouri, United

States of America. (2)Informatics Institute, University of Missouri, Columbia,

Missouri, United States of America; Department of Computer Science, University of

Missouri, Columbia, Missouri, United States of America. (3)Informatics Institute,

University of Missouri, Columbia, Missouri, United States of America; Department

of Computer Science, University of Missouri, Columbia, Missouri, United States of

America; Bond Life Science Center, University of Missouri, Columbia, Missouri,

United States of America.

Single nucleotide polymorphisms (SNPs) are among the most common types of genetic

variation in complex genetic disorders. A growing number of studies link the

functional role of SNPs with the networks and pathways mediated by the

disease-associated genes. For example, many non-synonymous missense SNPs (nsSNPs)

have been found near or inside the protein-protein interaction (PPI) interfaces.

Determining whether such nsSNP will disrupt or preserve a PPI is a challenging

task to address, both experimentally and computationally. Here, we present this

task as three related classification problems, and develop a new computational

method, called the SNP-IN tool (non-synonymous SNP INteraction effect predictor).

Our method predicts the effects of nsSNPs on PPIs, given the interaction's

structure. It leverages supervised and semi-supervised feature-based classifiers,

including our new Random Forest self-learning protocol. The classifiers are

trained based on a dataset of comprehensive mutagenesis studies for 151 PPI

complexes, with experimentally determined binding affinities of the mutant and

wild-type interactions. Three classification problems were considered: (1) a

2-class problem (strengthening/weakening PPI mutations), (2) another 2-class

problem (mutations that disrupt/preserve a PPI), and (3) a 3-class classification

(detrimental/neutral/beneficial mutation effects). In total, 11 different

supervised and semi-supervised classifiers were trained and assessed resulting in

a promising performance, with the weighted f-measure ranging from 0.87 for

Problem 1 to 0.70 for the most challenging Problem 3. By integrating prediction

results of the 2-class classifiers into the 3-class classifier, we further

improved its performance for Problem 3. To demonstrate the utility of SNP-IN

tool, it was applied to study the nsSNP-induced rewiring of two disease-centered

networks. The accurate and balanced performance of SNP-IN tool makes it readily

available to study the rewiring of large-scale protein-protein interaction

networks, and can be useful for functional annotation of disease-associated SNPs.

SNIP-IN tool is freely accessible as a web-server at

http://korkinlab.org/snpintool/.

DOI: 10.1371/journal.pcbi.1003592

PMCID: PMC4006705

PMID: 24784581 [Indexed for MEDLINE]

853. PLoS One. 2014 May 1;9(5):e94608. doi: 10.1371/journal.pone.0094608. eCollection

2014.

ModuleRole: a tool for modulization, role determination and visualization in

protein-protein interaction networks.

Li G(1), Li M(2), Zhang Y(3), Wang D(4), Li R(5), Guimerà R(6), Gao JT(1), Zhang

MQ(7).

Author information:

(1)MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for

Synthetic & Systems Biology, TNLIST, Department of Automation, Tsinghua

University, Beijing, People's Republic of China. (2)Bioinformatics Research

Group, Key Laboratory of Intelligent Information Processing, Advanced Computer

Research Center, Institute of Computing Technology, Chinese Academy of Sciences,

Beijing, People's Republic of China; University of Chinese Academy of Sciences,

Beijing, People's Republic of China. (3)Department of Basic Medical Sciences,

School of Medicine, Tsinghua University, Beijing, People's Republic of China.

(4)MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for

Synthetic & Systems Biology, TNLIST, Department of Automation, Tsinghua

University, Beijing, People's Republic of China; Department of Basic Medical

Sciences, School of Medicine, Tsinghua University, Beijing, People's Republic of

China. (5)Stowers Institute for Medical Research, Kansas City, Missouri, United

States of America; Department of Molecular and Integrative Physiology, University

of Kansas Medical Center, Kansas City, Kansas, United States of America.

(6)Institucit of Molecular and Cell Biology, Center for System, Barcelona,

Catalonia; Departament d'Enginyeria Qu Cell Biology, Center for SystemBarce,

Tarragona, Catalonia. (7)Department of Molecular and Cell Biology, Center for

Systems Biology, the University of Texas at Dallas, Richardson, Texas, United

States of America; MOE Key Laboratory of Bioinformatics; Bioinformatics Division

and Center for Synthetic & Systems Biology, TNLIST, Department of Automation,

Tsinghua University, Beijing, People's Republic of China.

Rapidly increasing amounts of (physical and genetic) protein-protein interaction

(PPI) data are produced by various high-throughput techniques, and interpretation

of these data remains a major challenge. In order to gain insight into the

organization and structure of the resultant large complex networks formed by

interacting molecules, using simulated annealing, a method based on the node

connectivity, we developed ModuleRole, a user-friendly web server tool which

finds modules in PPI network and defines the roles for every node, and produces

files for visualization in Cytoscape and Pajek. For given proteins, it analyzes

the PPI network from BioGRID database, finds and visualizes the modules these

proteins form, and then defines the role every node plays in this network, based

on two topological parameters Participation Coefficient and Z-score. This is the

first program which provides interactive and very friendly interface for

biologists to find and visualize modules and roles of proteins in PPI network. It

can be tested online at the website http://www.bioinfo.org/modulerole/index.php,

which is free and open to all users and there is no login requirement, with demo

data provided by "User Guide" in the menu Help. Non-server application of this

program is considered for high-throughput data with more than 200 nodes or user's

own interaction datasets. Users are able to bookmark the web link to the result

page and access at a later time. As an interactive and highly customizable

application, ModuleRole requires no expert knowledge in graph theory on the user

side and can be used in both Linux and Windows system, thus a very useful tool

for biologist to analyze and visualize PPI networks from databases such as

BioGRID.AVAILABILITY: ModuleRole is implemented in Java and C, and is freely

available at http://www.bioinfo.org/modulerole/index.php. Supplementary

information (user guide, demo data) is also available at this website. API for

ModuleRole used for this program can be obtained upon request.

DOI: 10.1371/journal.pone.0094608

PMCID: PMC4006751

PMID: 24788790 [Indexed for MEDLINE]

854. Proteins. 2014 May;82(5):794-814. doi: 10.1002/prot.24459. Epub 2013 Nov 22.

Princeton\_TIGRESS: protein geometry refinement using simulations and support

vector machines.

Khoury GA(1), Tamamis P, Pinnaduwage N, Smadbeck J, Kieslich CA, Floudas CA.

Author information:

(1)Department of Chemical and Biological Engineering, Princeton University,

Princeton, New Jersey, 08540.

Protein structure refinement aims to perform a set of operations given a

predicted structure to improve model quality and accuracy with respect to the

native in a blind fashion. Despite the numerous computational approaches to the

protein refinement problem reported in the previous three CASPs, an overwhelming

majority of methods degrade models rather than improve them. We initially

developed a method tested using blind predictions during CASP10 which was

officially ranked in 5th place among all methods in the refinement category.

Here, we present Princeton\_TIGRESS, which when benchmarked on all CASP 7,8,9, and

10 refinement targets, simultaneously increased GDT\_TS 76% of the time with an

average improvement of 0.83 GDT\_TS points per structure. The method was

additionally benchmarked on models produced by top performing three-dimensional

structure prediction servers during CASP10. The robustness of the

Princeton\_TIGRESS protocol was also tested for different random seeds. We make

the Princeton\_TIGRESS refinement protocol freely available as a web server at

http://atlas.princeton.edu/refinement. Using this protocol, one can consistently

refine a prediction to help bridge the gap between a predicted structure and the

actual native structure.

Copyright © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24459

PMID: 24174311 [Indexed for MEDLINE]

855. BMC Bioinformatics. 2014 Apr 27;15:119. doi: 10.1186/1471-2105-15-119.

ConSole: using modularity of contact maps to locate solenoid domains in protein

structures.

Hrabe T, Godzik A(1).

Author information:

(1)Program in Bioinformatics and Systems Biology, Sanford-Burnham Medical

Research Institute, 92037 La Jolla, CA, USA. adam@godziklab.org.

BACKGROUND: Periodic proteins, characterized by the presence of multiple repeats

of short motifs, form an interesting and seldom-studied group. Due to often

extreme divergence in sequence, detection and analysis of such motifs is

performed more reliably on the structural level. Yet, few algorithms have been

developed for the detection and analysis of structures of periodic proteins.

RESULTS: ConSole recognizes modularity in protein contact maps, allowing for

precise identification of repeats in solenoid protein structures, an important

subgroup of periodic proteins. Tests on benchmarks show that ConSole has higher

recognition accuracy as compared to Raphael, the only other publicly available

solenoid structure detection tool. As a next step of ConSole analysis, we show

how detection of solenoid repeats in structures can be used to improve sequence

recognition of these motifs and to detect subtle irregularities of repeat lengths

in three solenoid protein families.

CONCLUSIONS: The ConSole algorithm provides a fast and accurate tool to recognize

solenoid protein structures as a whole and to identify individual solenoid repeat

units from a structure. ConSole is available as a web-based, interactive server

and is available for download at http://console.sanfordburnham.org.

DOI: 10.1186/1471-2105-15-119

PMCID: PMC4021314

PMID: 24766872 [Indexed for MEDLINE]

856. J Mol Biol. 2014 Apr 17;426(8):1861-9. doi: 10.1016/j.jmb.2014.02.003. Epub 2014

Feb 9.

NP-Sticky: a web server for optimizing DNA ligation with non-palindromic sticky

ends.

Ng DT(1), Sarkar CA(2).

Author information:

(1)Department of Bioengineering, University of Pennsylvania, Philadelphia, PA

19104, USA. (2)Department of Biomedical Engineering, University of Minnesota,

Minneapolis, MN 55455, USA. Electronic address: csarkar@umn.edu.

High-efficiency DNA ligation is vital for many molecular biology experiments, and

it is best achieved using reactants with non-palindromic sticky ends to maximize

specificity. However, optimizing such multi-parametric ligation reactions often

involves extensive trial and error. We have developed a freely available

Web-based ligation calculator, NP-Sticky (http://sarkarlab.umn.edu/npsticky/),

that predicts product distribution for given reactant concentrations, thus

enabling straightforward computational optimization of these reactions. Built-in

schemes include two-piece and three-piece linear ligation, as well as

insert-vector circular ligation. The only parameters needed for the underlying

thermodynamic model are the free energies of ligation for each sticky end, which

can be estimated by the calculator from the overhang sequences or provided by the

user from direct experimental measurement. Free energies of sticky-end mismatches

are also calculated for determining the extent of byproduct formation. This

ligation calculator allows rapid identification of the optimal conditions for

maximizing incorporation, efficiency, and/or accuracy, based on specific needs.

Copyright © 2014 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jmb.2014.02.003

PMID: 24518657 [Indexed for MEDLINE]

857. Bioinformatics. 2014 Apr 15;30(8):1120-1128. Epub 2014 Jan 7.

Allerdictor: fast allergen prediction using text classification techniques.

Dang HX(1), Lawrence CB(2).

Author information:

(1)Virginia Bioinformatics Institute and Department of Biological Sciences,

Virginia Tech, Blacksburg, VA 24061, USA. (2)Virginia Bioinformatics Institute

and Department of Biological Sciences, Virginia Tech, Blacksburg, VA 24061, USA

Virginia Bioinformatics Institute and Department of Biological Sciences, Virginia

Tech, Blacksburg, VA 24061, USA.

MOTIVATION: Accurately identifying and eliminating allergens from

biotechnology-derived products are important for human health. From a biomedical

research perspective, it is also important to identify allergens in sequenced

genomes. Many allergen prediction tools have been developed during the past

years. Although these tools have achieved certain levels of specificity, when

applied to large-scale allergen discovery (e.g. at a whole-genome scale), they

still yield many false positives and thus low precision (even at low recall) due

to the extreme skewness of the data (allergens are rare). Moreover, the most

accurate tools are relatively slow because they use protein sequence alignment to

build feature vectors for allergen classifiers. Additionally, only web server

implementations of the current allergen prediction tools are publicly available

and are without the capability of large batch submission. These weaknesses make

large-scale allergen discovery ineffective and inefficient in the public domain.

RESULTS: We developed Allerdictor, a fast and accurate sequence-based allergen

prediction tool that models protein sequences as text documents and uses support

vector machine in text classification for allergen prediction. Test results on

multiple highly skewed datasets demonstrated that Allerdictor predicted allergens

with high precision over high recall at fast speed. For example, Allerdictor only

took ∼6 min on a single core PC to scan a whole Swiss-Prot database of ∼540 000

sequences and identified <1% of them as allergens.

AVAILABILITY AND IMPLEMENTATION: Allerdictor is implemented in Python and

available as standalone and web server versions at http://allerdictor.vbi.vt.edu

CONTACT: lawrence@vbi.vt.edu Supplementary information: Supplementary data are

available at Bioinformatics online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btu004

PMCID: PMC3982160

PMID: 24403538

858. Bioinformatics. 2014 Apr 15;30(8):1190-1192. Epub 2014 Jan 2.

CMGRN: a web server for constructing multilevel gene regulatory networks using

ChIP-seq and gene expression data.

Guan D(1), Shao J(1), Deng Y(1), Wang P(1), Zhao Z(1), Liang Y(1), Wang J(2), Yan

B(2).

Author information:

(1)Department of Biology, Hong Kong Baptist University, Hong Kong, China,

Department of Internal Medicine and Biochemistry, Rush University Medical Center,

Chicago, IL, USA, Department of Biochemistry and HKU-SIRI, Centre for Genomic

Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China

and Key Laboratory of Acupuncture and Medicine Research of Minister of Education

, Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210046, China.

(2)Department of Biology, Hong Kong Baptist University, Hong Kong, China,

Department of Internal Medicine and Biochemistry, Rush University Medical Center,

Chicago, IL, USA, Department of Biochemistry and HKU-SIRI, Centre for Genomic

Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China

and Key Laboratory of Acupuncture and Medicine Research of Minister of Education

, Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210046, China

Department of Biology, Hong Kong Baptist University, Hong Kong, China, Department

of Internal Medicine and Biochemistry, Rush University Medical Center, Chicago,

IL, USA, Department of Biochemistry and HKU-SIRI, Centre for Genomic Sciences,

LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China and Key

Laboratory of Acupuncture and Medicine Research of Minister of Education ,

Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210046, China.

ChIP-seq technology provides an accurate characterization of transcription or

epigenetic factors binding on genomic sequences. With integration of such

ChIP-based and other high-throughput information, it would be dedicated to

dissecting cross-interactions among multilevel regulators, genes and biological

functions. Here, we devised an integrative web server CMGRN (constructing

multilevel gene regulatory networks), to unravel hierarchical interactive

networks at different regulatory levels. The newly developed method used the

Bayesian network modeling to infer causal interrelationships among transcription

factors or epigenetic modifications by using ChIP-seq data. Moreover, it used

Bayesian hierarchical model with Gibbs sampling to incorporate binding signals of

these regulators and gene expression profile together for reconstructing gene

regulatory networks. The example applications indicate that CMGRN provides an

effective web-based framework that is able to integrate heterogeneous

high-throughput data and to reveal hierarchical 'regulome' and the associated

gene expression programs.AVAILABILITY: http://bioinfo.icts.hkbu.edu.hk/cmgrn;

http://www.byanbioinfo.org/cmgrn CONTACT: yanbinai6017@gmail.com or junwen@hku.hk

Supplementary Information: Supplementary data are available at Bioinformatics

online.

© The Author 2014. Published by Oxford University Press. All rights reserved. For

Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/bioinformatics/btt761

PMID: 24389658

859. BMC Bioinformatics. 2014 Apr 15;15:110. doi: 10.1186/1471-2105-15-110.

On finding bicliques in bipartite graphs: a novel algorithm and its application

to the integration of diverse biological data types.

Zhang Y, Phillips CA, Rogers GL, Baker EJ, Chesler EJ, Langston MA(1).

Author information:

(1)Department of Electrical Engineering and Computer Science, University of

Tennessee, Knoxville, TN 37996, USA. langston@eecs.utk.edu.

BACKGROUND: Integrating and analyzing heterogeneous genome-scale data is a huge

algorithmic challenge for modern systems biology. Bipartite graphs can be useful

for representing relationships across pairs of disparate data types, with the

interpretation of these relationships accomplished through an enumeration of

maximal bicliques. Most previously-known techniques are generally ill-suited to

this foundational task, because they are relatively inefficient and without

effective scaling. In this paper, a powerful new algorithm is described that

produces all maximal bicliques in a bipartite graph. Unlike most previous

approaches, the new method neither places undue restrictions on its input nor

inflates the problem size. Efficiency is achieved through an innovative

exploitation of bipartite graph structure, and through computational reductions

that rapidly eliminate non-maximal candidates from the search space. An iterative

selection of vertices for consideration based on non-decreasing common

neighborhood sizes boosts efficiency and leads to more balanced recursion trees.

RESULTS: The new technique is implemented and compared to previously published

approaches from graph theory and data mining. Formal time and space bounds are

derived. Experiments are performed on both random graphs and graphs constructed

from functional genomics data. It is shown that the new method substantially

outperforms the best previous alternatives.

CONCLUSIONS: The new method is streamlined, efficient, and particularly

well-suited to the study of huge and diverse biological data. A robust

implementation has been incorporated into GeneWeaver, an online tool for

integrating and analyzing functional genomics experiments, available at

http://geneweaver.org. The enormous increase in scalability it provides empowers

users to study complex and previously unassailable gene-set associations between

genes and their biological functions in a hierarchical fashion and on a

genome-wide scale. This practical computational resource is adaptable to almost

any applications environment in which bipartite graphs can be used to model

relationships between pairs of heterogeneous entities.

DOI: 10.1186/1471-2105-15-110

PMCID: PMC4038116

PMID: 24731198 [Indexed for MEDLINE]

860. PLoS One. 2014 Apr 15;9(4):e93907. doi: 10.1371/journal.pone.0093907. eCollection

2014.

MP3: a software tool for the prediction of pathogenic proteins in genomic and

metagenomic data.

Gupta A(1), Kapil R(1), Dhakan DB(1), Sharma VK(1).

Author information:

(1)MetaInformatics Laboratory, Metagenomics and Systems Biology Group, Department

of Biological Sciences, Indian Institute of Science Education and Research

Bhopal, Madhya Pradesh, India.

The identification of virulent proteins in any de-novo sequenced genome is useful

in estimating its pathogenic ability and understanding the mechanism of

pathogenesis. Similarly, the identification of such proteins could be valuable in

comparing the metagenome of healthy and diseased individuals and estimating the

proportion of pathogenic species. However, the common challenge in both the above

tasks is the identification of virulent proteins since a significant proportion

of genomic and metagenomic proteins are novel and yet unannotated. The currently

available tools which carry out the identification of virulent proteins provide

limited accuracy and cannot be used on large datasets. Therefore, we have

developed an MP3 standalone tool and web server for the prediction of pathogenic

proteins in both genomic and metagenomic datasets. MP3 is developed using an

integrated Support Vector Machine (SVM) and Hidden Markov Model (HMM) approach to

carry out highly fast, sensitive and accurate prediction of pathogenic proteins.

It displayed Sensitivity, Specificity, MCC and accuracy values of 92%, 100%, 0.92

and 96%, respectively, on blind dataset constructed using complete proteins. On

the two metagenomic blind datasets (Blind A: 51-100 amino acids and Blind B:

30-50 amino acids), it displayed Sensitivity, Specificity, MCC and accuracy

values of 82.39%, 97.86%, 0.80 and 89.32% for Blind A and 71.60%, 94.48%, 0.67

and 81.86% for Blind B, respectively. In addition, the performance of MP3 was

validated on selected bacterial genomic and real metagenomic datasets. To our

knowledge, MP3 is the only program that specializes in fast and accurate

identification of partial pathogenic proteins predicted from short (100-150 bp)

metagenomic reads and also performs exceptionally well on complete protein

sequences. MP3 is publicly available at

http://metagenomics.iiserb.ac.in/mp3/index.php.

DOI: 10.1371/journal.pone.0093907

PMCID: PMC3988012

PMID: 24736651 [Indexed for MEDLINE]

861. Gene. 2014 Apr 10;539(1):152-3. doi: 10.1016/j.gene.2014.02.007. Epub 2014 Feb

11.

An intuitive graphical webserver for multiple-choice protein sequence search.

Banky D(1), Szalkai B(2), Grolmusz V(3).

Author information:

(1)Protein Information Technology Group, Eotvos University, Pazmany Peter stny.

1/C, H-1117 Budapest, Hungary; Uratim Ltd., H-1118 Budapest, Hungary. Electronic

address: banky@pitgroup.org. (2)Protein Information Technology Group, Eotvos

University, Pazmany Peter stny. 1/C, H-1117 Budapest, Hungary. Electronic

address: szalkai@pitgroup.org. (3)Protein Information Technology Group, Eotvos

University, Pazmany Peter stny. 1/C, H-1117 Budapest, Hungary; Uratim Ltd.,

H-1118 Budapest, Hungary. Electronic address: grolmusz@pitgroup.org.

Every day tens of thousands of sequence searches and sequence alignment queries

are submitted to webservers. The capitalized word "BLAST" becomes a verb,

describing the act of performing sequence search and alignment. However, if one

needs to search for sequences that contain, for example, two hydrophobic and

three polar residues at five given positions, the query formation on the most

frequently used webservers will be difficult. Some servers support the formation

of queries with regular expressions, but most of the users are unfamiliar with

their syntax. Here we present an intuitive, easily applicable webserver, the

Protein Sequence Analysis server, that allows the formation of multiple choice

queries by simply drawing the residues to their positions; if more than one

residue are drawn to the same position, then they will be nicely stacked on the

user interface, indicating the multiple choice at the given position. This

computer-game-like interface is natural and intuitive, and the coloring of the

residues makes possible to form queries requiring not just certain amino acids in

the given positions, but also small nonpolar, negatively charged, hydrophobic,

positively charged, or polar ones. The webserver is available at

http://psa.pitgroup.org.

Copyright © 2014 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.gene.2014.02.007

PMID: 24525401 [Indexed for MEDLINE]

862. PLoS One. 2014 Apr 8;9(4):e94088. doi: 10.1371/journal.pone.0094088. eCollection

2014.

NEMiD: a web-based curated microbial diversity database with geo-based plotting.

Bhattacharjee K(1), Joshi SR(1).

Author information:

(1)Microbiology Laboratory, Department of Biotechnology and Bioinformatics,

North-Eastern Hill University, Shillong, Meghalaya, India.

The majority of the Earth's microbes remain unknown, and that their potential

utility cannot be exploited until they are discovered and characterized. They

provide wide scope for the development of new strains as well as biotechnological

uses. The documentation and bioprospection of microorganisms carry enormous

significance considering their relevance to human welfare. This calls for an

urgent need to develop a database with emphasis on the microbial diversity of the

largest untapped reservoirs in the biosphere. The data annotated in the

North-East India Microbial database (NEMiD) were obtained by the isolation and

characterization of microbes from different parts of the Eastern Himalayan

region. The database was constructed as a relational database management system

(RDBMS) for data storage in MySQL in the back-end on a Linux server and

implemented in an Apache/PHP environment. This database provides a base for

understanding the soil microbial diversity pattern in this megabiodiversity

hotspot and indicates the distribution patterns of various organisms along with

identification. The NEMiD database is freely available at

www.mblabnehu.info/nemid/.

DOI: 10.1371/journal.pone.0094088

PMCID: PMC3979743

PMID: 24714636 [Indexed for MEDLINE]

863. BMC Bioinformatics. 2014 Apr 3;15:96. doi: 10.1186/1471-2105-15-96.

Quantum coupled mutation finder: predicting functionally or structurally

important sites in proteins using quantum Jensen-Shannon divergence and CUDA

programming.

Gültas M(1), Düzgün G, Herzog S, Jäger SJ, Meckbach C, Wingender E, Waack S.

Author information:

(1)Institute of Computer Science, University of Göttingen, Goldschmidtstr, 7,

37077 Göttingen, Germany. waack@cs.uni-goettingen.de.

BACKGROUND: The identification of functionally or structurally important

non-conserved residue sites in protein MSAs is an important challenge for

understanding the structural basis and molecular mechanism of protein functions.

Despite the rich literature on compensatory mutations as well as sequence

conservation analysis for the detection of those important residues, previous

methods often rely on classical information-theoretic measures. However, these

measures usually do not take into account dis/similarities of amino acids which

are likely to be crucial for those residues. In this study, we present a new

method, the Quantum Coupled Mutation Finder (QCMF) that incorporates significant

dis/similar amino acid pair signals in the prediction of functionally or

structurally important sites.

RESULTS: The result of this study is twofold. First, using the essential sites of

two human proteins, namely epidermal growth factor receptor (EGFR) and

glucokinase (GCK), we tested the QCMF-method. The QCMF includes two metrics based

on quantum Jensen-Shannon divergence to measure both sequence conservation and

compensatory mutations. We found that the QCMF reaches an improved performance in

identifying essential sites from MSAs of both proteins with a significantly

higher Matthews correlation coefficient (MCC) value in comparison to previous

methods. Second, using a data set of 153 proteins, we made a pairwise comparison

between QCMF and three conventional methods. This comparison study strongly

suggests that QCMF complements the conventional methods for the identification of

correlated mutations in MSAs.

CONCLUSIONS: QCMF utilizes the notion of entanglement, which is a major resource

of quantum information, to model significant dissimilar and similar amino acid

pair signals in the detection of functionally or structurally important sites.

Our results suggest that on the one hand QCMF significantly outperforms the

previous method, which mainly focuses on dissimilar amino acid signals, to detect

essential sites in proteins. On the other hand, it is complementary to the

existing methods for the identification of correlated mutations. The method of

QCMF is computationally intensive. To ensure a feasible computation time of the

QCMF's algorithm, we leveraged Compute Unified Device Architecture (CUDA).The

QCMF server is freely accessible at http://qcmf.informatik.uni-goettingen.de/.

DOI: 10.1186/1471-2105-15-96

PMCID: PMC4098773

PMID: 24694117 [Indexed for MEDLINE]

864. BMC Genomics. 2014 Apr 3;15:256. doi: 10.1186/1471-2164-15-256.

HGCS: an online tool for prioritizing disease-causing gene variants by biological

distance.

Itan Y(1), Mazel M, Mazel B, Abhyankar A, Nitschke P, Quintana-Murci L,

Boisson-Dupuis S, Boisson B, Abel L, Zhang SY, Casanova JL.

Author information:

(1)St, Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller

Branch, The Rockefeller University, New York, NY, USA. yitan@rockefeller.edu.

BACKGROUND: Identifying the genotypes underlying human disease phenotypes is a

fundamental step in human genetics and medicine. High-throughput genomic

technologies provide thousands of genetic variants per individual. The causal

genes of a specific phenotype are usually expected to be functionally close to

each other. According to this hypothesis, candidate genes are picked from

high-throughput data on the basis of their biological proximity to core genes -

genes already known to be responsible for the phenotype. There is currently no

effective gene-centric online interface for this purpose.

RESULTS: We describe here the human gene connectome server (HGCS), a powerful,

easy-to-use interactive online tool enabling researchers to prioritize any list

of genes according to their biological proximity to core genes associated with

the phenotype of interest. We also make available an updated and extended version

for all human gene-specific connectomes. The HGCS is freely available to

noncommercial users from: http://hgc.rockefeller.edu.

CONCLUSIONS: The HGCS should help investigators from diverse fields to identify

new disease-causing candidate genes more effectively, via a user-friendly online

interface.

DOI: 10.1186/1471-2164-15-256

PMCID: PMC4051124

PMID: 24694260 [Indexed for MEDLINE]

865. Bioinformatics. 2014 Apr 1;30(7):1013-4. doi: 10.1093/bioinformatics/btt655. Epub

2013 Nov 9.

AVIA: an interactive web-server for annotation, visualization and impact analysis

of genomic variations.

Vuong H(1), Stephens RM, Volfovsky N.

Author information:

(1)Advanced Biomedical Computing Center (ABCC), Information Systems Program,

SAIC-Frederick Inc., Frederick National Laboratory for Cancer Research,

Frederick, MD 21702, USA.

MOTIVATION: The plethora of information that emerges from large-scale genome

characterization studies has triggered the development of computational

frameworks and tools for efficient analysis, interpretation and visualization of

genomic data. Functional annotation of genomic variations and the ability to

visualize the data in the context of whole genome and/or multiple genomes has

remained a challenging task. We have developed an interactive web-based tool,

AVIA (Annotation, Visualization and Impact Analysis), to explore and interpret

large sets of genomic variations (single nucleotide variations and

insertion/deletions) and to help guide and summarize genomic experiments. The

annotation, summary plots and tables are packaged and can be downloaded by the

user from the email link provided.

AVAILABILITY AND IMPLEMENTATION: http://avia.abcc.ncifcrf.gov.

DOI: 10.1093/bioinformatics/btt655

PMCID: PMC3967104

PMID: 24215028 [Indexed for MEDLINE]

866. BMC Syst Biol. 2014 Apr 1;8:40. doi: 10.1186/1752-0509-8-40.

NatalieQ: a web server for protein-protein interaction network querying.

El-Kebir M(1), Brandt BW, Heringa J, Klau GW.

Author information:

(1)Life Sciences, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG

Amsterdam, the Netherlands. m.el-kebir@cwi.nl.

BACKGROUND: Molecular interactions need to be taken into account to adequately

model the complex behavior of biological systems. These interactions are captured

by various types of biological networks, such as metabolic, gene-regulatory,

signal transduction and protein-protein interaction networks. We recently

developed Natalie, which computes high-quality network alignments via advanced

methods from combinatorial optimization.

RESULTS: Here, we present NatalieQ, a web server for topology-based alignment of

a specified query protein-protein interaction network to a selected target

network using the Natalie algorithm. By incorporating similarity at both the

sequence and the network level, we compute alignments that allow for the transfer

of functional annotation as well as for the prediction of missing interactions.

We illustrate the capabilities of NatalieQ with a biological case study involving

the Wnt signaling pathway.

CONCLUSIONS: We show that topology-based network alignment can produce results

complementary to those obtained by using sequence similarity alone. We also

demonstrate that NatalieQ is able to predict putative interactions. The server is

available at: http://www.ibi.vu.nl/programs/natalieq/.

DOI: 10.1186/1752-0509-8-40

PMCID: PMC3998945

PMID: 24690407 [Indexed for MEDLINE]

867. Curr Genomics. 2014 Apr;15(2):113-21. doi: 10.2174/1389202915999140328163125.

Identification of Horizontally-transferred Genomic Islands and Genome

Segmentation Points by Using the GC Profile Method.

Zhang R(1), Ou HY(2), Gao F(3), Luo H(3).

Author information:

(1)Center for Molecular Medicine and Genetics, School of Medicine, Wayne State

University, Detroit, MI, USA; (2)State Key Laboratory of Microbial Metabolism and

School of Life Sciences & Biotechnology, Shanghai Jiaotong University, Shanghai

200030, China; (3)Department of Physics, Tianjin University, Tianjin, 300072,

China.

The nucleotide composition of genomes undergoes dramatic variations among all

three kingdoms of life. GC content, an important characteristic for a genome, is

related to many important functions, and therefore GC content and its

distribution are routinely reported for sequenced genomes. Traditionally, GC

content distribution is assessed by computing GC contents in windows that slide

along the genome. Disadvantages of this routinely used window-based method

include low resolution and low sensitivity. Additionally, different window sizes

result in different GC content distribution patterns within the same genome. We

proposed a windowless method, the GC profile, for displaying GC content

variations across the genome. Compared to the window-based method, the GC profile

has the following advantages: 1) higher sensitivity, because of

variation-amplifying procedures; 2) higher resolution, because boundaries between

domains can be determined at one single base pair; 3) uniqueness, because the GC

profile is unique for a given genome and 4) the capacity to show both global and

regional GC content distributions. These characteristics are useful in

identifying horizontally-transferred genomic islands and homogenous GC-content

domains. Here, we review the applications of the GC profile in identifying

genomic islands and genome segmentation points, and in serving as a platform to

integrate with other algorithms for genome analysis. A web server generating GC

profiles and implementing relevant genome segmentation algorithms is available

at: www.zcurve.net.

DOI: 10.2174/1389202915999140328163125

PMCID: PMC4009839

PMID: 24822029

868. J Biomol Struct Dyn. 2014 Apr;32(4):661-8. doi: 10.1080/07391102.2013.787026.

Epub 2013 May 10.

mulPBA: an efficient multiple protein structure alignment method based on a

structural alphabet.

Léonard S(1), Joseph AP, Srinivasan N, Gelly JC, de Brevern AG.

Author information:

(1)a INSERM UMR-S 665, DSIMB , 6, rue Alexandre Cabanel, F-75739 , Paris , France

.

The increasing number of available protein structures requires efficient tools

for multiple structure comparison. Indeed, multiple structural alignments are

essential for the analysis of function, evolution and architecture of protein

structures. For this purpose, we proposed a new web server called multiple

Protein Block Alignment (mulPBA). This server implements a method based on a

structural alphabet to describe the backbone conformation of a protein chain in

terms of dihedral angles. This 'sequence-like' representation enables the use of

powerful sequence alignment methods for primary structure comparison, followed by

an iterative refinement of the structural superposition. This approach yields

alignments superior to most of the rigid-body alignment methods and highly

comparable with the flexible structure comparison approaches. We implement this

method in a web server designed to do multiple structure superimpositions from a

set of structures given by the user. Outputs are given as both sequence alignment

and superposed 3D structures visualized directly by static images generated by

PyMol or through a Jmol applet allowing dynamic interaction. Multiple global

quality measures are given. Relatedness between structures is indicated by a

distance dendogram. Superimposed structures in PDB format can be also downloaded,

and the results are quickly obtained. mulPBA server can be accessed at

www.dsimb.inserm.fr/dsimb\_tools/mulpba/ .

DOI: 10.1080/07391102.2013.787026

PMID: 23659291 [Indexed for MEDLINE]

869. Mol Inform. 2014 Apr;33(4):276-85. doi: 10.1002/minf.201300027. Epub 2014 Mar 18.

LECTINPred: web Server that Uses Complex Networks of Protein Structure for

Prediction of Lectins with Potential Use as Cancer Biomarkers or in Parasite

Vaccine Design.

Munteanu CR(1), Pedreira N(2), Dorado J(2), Pazos A(2), Pérez-Montoto LG(3),

Ubeira FM(3), González-Díaz H(4,)(5).

Author information:

(1)Department of Information and Communication Technologies, Computer Science

Faculty, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

phone: (+34) 981167000, ext. 1302; fax: (+34) 981167160. crm.publish@gmail.com.

(2)Department of Information and Communication Technologies, Computer Science

Faculty, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

phone: (+34) 981167000, ext. 1302; fax: (+34) 981167160. (3)Department of

Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de

Compostela, 15782 Santiago de Compostela, Spain. (4)Department of Organic

Chemistry II, Faculty of Science and Technology, University of the Basque

Country, UPV/EHU, 48940, Leioa, Bizkaia, Spain. (5)IKERBASQUE, Basque Foundation

for Science, 48011, Bilbao, Spain.

Lectins (Ls) play an important role in many diseases such as different types of

cancer, parasitic infections and other diseases. Interestingly, the Protein Data

Bank (PDB) contains +3000 protein 3D structures with unknown function. Thus, we

can in principle, discover new Ls mining non-annotated structures from PDB or

other sources. However, there are no general models to predict new biologically

relevant Ls based on 3D chemical structures. We used the MARCH-INSIDE software to

calculate the Markov-Shannon 3D electrostatic entropy parameters for the complex

networks of protein structure of 2200 different protein 3D structures, including

1200 Ls. We have performed a Linear Discriminant Analysis (LDA) using these

parameters as inputs in order to seek a new Quantitative Structure-Activity

Relationship (QSAR) model, which is able to discriminate 3D structure of Ls from

other proteins. We implemented this predictor in the web server named LECTINPred,

freely available at http://bio-aims.udc.es/LECTINPred.php. This web server showed

the following goodness-of-fit statistics: Sensitivity=96.7 % (for Ls),

Specificity=87.6 % (non-active proteins), and Accuracy=92.5 % (for all proteins),

considering altogether both the training and external prediction series. In mode

2, users can carry out an automatic retrieval of protein structures from PDB. We

illustrated the use of this server, in operation mode 1, performing a data mining

of PDB. We predicted Ls scores for +2000 proteins with unknown function and

selected the top-scored ones as possible lectins. In operation mode 2, LECTINPred

can also upload 3D structural models generated with structure-prediction tools

like LOMETS or PHYRE2. The new Ls are expected to be of relevance as cancer

biomarkers or useful in parasite vaccine design.

© 2014 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

DOI: 10.1002/minf.201300027

PMID: 27485774

870. BMC Bioinformatics. 2014 Mar 31;15:93. doi: 10.1186/1471-2105-15-93.

SPiCE: a web-based tool for sequence-based protein classification and

exploration.

van den Berg BA(1), Reinders MJ, Roubos JA, de Ridder D.

Author information:

(1)Delft Bioinformatics Lab, Department of Intelligent Systems, Faculty

Electrical Engineering, Mathematics and Computer Science, Delft University of

Technology, Mekelweg 4, 2628CD, Delft, The Netherlands.

b.a.vandenberg@tudelft.nl.

BACKGROUND: Amino acid sequences and features extracted from such sequences have

been used to predict many protein properties, such as subcellular localization or

solubility, using classifier algorithms. Although software tools are available

for both feature extraction and classifier construction, their application is not

straightforward, requiring users to install various packages and to convert data

into different formats. This lack of easily accessible software hampers quick,

explorative use of sequence-based classification techniques by biologists.

RESULTS: We have developed the web-based software tool SPiCE for exploring

sequence-based features of proteins in predefined classes. It offers data

upload/download, sequence-based feature calculation, data visualization and

protein classifier construction and testing in a single integrated, interactive

environment. To illustrate its use, two example datasets are included showing the

identification of differences in amino acid composition between proteins yielding

low and high production levels in fungi and low and high expression levels in

yeast, respectively.

CONCLUSIONS: SPiCE is an easy-to-use online tool for extracting and exploring

sequence-based features of sets of proteins, allowing non-experts to apply

advanced classification techniques. The tool is available at

http://helix.ewi.tudelft.nl/spice.

DOI: 10.1186/1471-2105-15-93

PMCID: PMC4021553

PMID: 24685258 [Indexed for MEDLINE]

871. J Clin Bioinforma. 2014 Mar 31;4(1):5. doi: 10.1186/2043-9113-4-5.

FISH Oracle 2: a web server for integrative visualization of genomic data in

cancer research.

Mader M, Simon R, Kurtz S(1).

Author information:

(1)Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146

Hamburg, Germany. kurtz@zbh.uni-hamburg.de.

BACKGROUND: A comprehensive view on all relevant genomic data is instrumental for

understanding the complex patterns of molecular alterations typically found in

cancer cells. One of the most effective ways to rapidly obtain an overview of

genomic alterations in large amounts of genomic data is the integrative

visualization of genomic events.

RESULTS: We developed FISH Oracle 2, a web server for the interactive

visualization of different kinds of downstream processed genomics data typically

available in cancer research. A powerful search interface and a fast

visualization engine provide a highly interactive visualization for such data.

High quality image export enables the life scientist to easily communicate their

results. A comprehensive data administration allows to keep track of the

available data sets. We applied FISH Oracle 2 to published data and found

evidence that, in colorectal cancer cells, the gene TTC28 may be inactivated in

two different ways, a fact that has not been published before.

CONCLUSIONS: The interactive nature of FISH Oracle 2 and the possibility to

store, select and visualize large amounts of downstream processed data support

life scientists in generating hypotheses. The export of high quality images

supports explanatory data visualization, simplifying the communication of new

biological findings. A FISH Oracle 2 demo server and the software is available at

http://www.zbh.uni-hamburg.de/fishoracle.

DOI: 10.1186/2043-9113-4-5

PMCID: PMC4230720

PMID: 24684958

872. Sci Rep. 2014 Mar 26;4:4474. doi: 10.1038/srep04474.

GeneSense: a new approach for human gene annotation integrated with

protein-protein interaction networks.

Chen Z(1), Zhang T(2), Lin J(3), Yan Z(4), Wang Y(5), Zheng W(6), Weng KC(7).

Author information:

(1)BioMedSense Laboratory, Shanghai 200433, China. (2)Department of

Otorhinolaryngology, Head and Neck Surgery, The First Affiliated Hospital, Harbin

Medical University, Harbin 150081, China. (3)Department of Biotechnology, Guilin

Medical University, Guilin 541004, China. (4)Shanghai Yingyun Biotech Inc.,

Shanghai 200433, China. (5)College of Basic Science, Zhejiang Chinese Medical

University, Hangzhou 310053, China. (6)Department of Pathology, Changhai

Hospital, Shanghai 200433, China. (7)Nplex Laboratory, San Jose, CA 95134, USA.

Virtually all cellular functions involve protein-protein interactions (PPIs). As

an increasing number of PPIs are identified and vast amount of information

accumulated, researchers are finding different ways to interrogate the data and

understand the interactions in context. However, it is widely recognized that a

significant portion of the data is scattered, redundant, not considered high

quality, and not readily accessible to researchers in a systematic fashion. In

addition, it is challenging to identify the optimal protein targets in the

current PPI networks. The GeneSense server was developed to integrate gene

annotation and PPI networks in an expandable architecture that incorporates

selected databases with the aim to assemble, analyze, evaluate and disseminate

protein-protein association information in a comprehensive and user-friendly

manner. Three network models including nodenet, leafnet and loopnet are used to

identify the optimal protein targets in the complex networks. GeneSense is freely

available at www.biomedsense.org/genesense.php.

DOI: 10.1038/srep04474

PMCID: PMC3966033

PMID: 24667292 [Indexed for MEDLINE]

873. ACS Synth Biol. 2014 Mar 21;3(3):192-6. doi: 10.1021/sb400178c. Epub 2014 Jan 6.

Database construction for PromoterCAD: synthetic promoter design for mammals and

plants.

Nishikata K(1), Cox RS 3rd, Shimoyama S, Yoshida Y, Matsui M, Makita Y, Toyoda T.

Author information:

(1)Integrated Database Unit, Advanced Center for Computing and Communication

(ACCC), RIKEN , 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.

Synthetic promoters can control a gene's timing, location, and expression level.

The PromoterCAD web server ( http://promotercad.org ) allows the design of

synthetic promoters to control plant gene expression, by novel arrangement of

cis-regulatory elements. Recently, we have expanded PromoterCAD's scope with

additional plant and animal data: (1) PLACE (Plant Cis-acting Regulatory DNA

Elements), including various sized sequence motifs; (2) PEDB (Mammalian

Promoter/Enhancer Database), including gene expression data for mammalian

tissues. The plant PromoterCAD data now contains 22 000 Arabidopsis thaliana

genes, 2 200 000 microarray measurements in 20 growth conditions and 79 tissue

organs and developmental stages, while the new mammalian PromoterCAD data

contains 679 Mus musculus genes and 65 000 microarray measurements in 96 tissue

organs and cell types ( http://promotercad.org/mammal/ ). This work presents

step-by-step instructions for adding both regulatory motif and gene expression

data to PromoterCAD, to illustrate how users can expand PromoterCAD functionality

for their own applications and organisms.

DOI: 10.1021/sb400178c

PMID: 24364365 [Indexed for MEDLINE]

874. Int J Mol Sci. 2014 Mar 19;15(3):4915-37. doi: 10.3390/ijms15034915.

iNR-Drug: predicting the interaction of drugs with nuclear receptors in cellular

networking.

Fan YN(1), Xiao X(2), Min JL(3), Chou KC(4).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen 333046,

Jiangxi, China. yuenong.f@163.com. (2)Computer Department, Jing-De-Zhen Ceramic

Institute, Jingdezhen 333046, Jiangxi, China. xxiao@gordonlifescience.org.

(3)Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen 333046,

Jiangxi, China. minjianliang@126.com. (4)Center of Excellence in Genomic Medicine

Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia.

kcchou@gordonlifescience.org.

Nuclear receptors (NRs) are closely associated with various major diseases such

as cancer, diabetes, inflammatory disease, and osteoporosis. Therefore, NRs have

become a frequent target for drug development. During the process of developing

drugs against these diseases by targeting NRs, we are often facing a problem:

Given a NR and chemical compound, can we identify whether they are really in

interaction with each other in a cell? To address this problem, a predictor

called "iNR-Drug" was developed. In the predictor, the drug compound concerned

was formulated by a 256-D (dimensional) vector derived from its molecular

fingerprint, and the NR by a 500-D vector formed by incorporating its sequential

evolution information and physicochemical features into the general form of

pseudo amino acid composition, and the prediction engine was operated by the SVM

(support vector machine) algorithm. Compared with the existing prediction methods

in this area, iNR-Drug not only can yield a higher success rate, but is also

featured by a user-friendly web-server established at

http://www.jci-bioinfo.cn/iNR-Drug/, which is particularly useful for most

experimental scientists to obtain their desired data in a timely manner. It is

anticipated that the iNR-Drug server may become a useful high throughput tool for

both basic research and drug development, and that the current approach may be

easily extended to study the interactions of drug with other targets as well.

DOI: 10.3390/ijms15034915

PMCID: PMC3975431

PMID: 24651462 [Indexed for MEDLINE]

875. PLoS One. 2014 Mar 19;9(3):e89545. doi: 10.1371/journal.pone.0089545. eCollection

2014.

HybridGO-Loc: mining hybrid features on gene ontology for predicting subcellular

localization of multi-location proteins.

Wan S(1), Mak MW(1), Kung SY(2).

Author information:

(1)Department of Electronic and Information Engineering, The Hong Kong

Polytechnic University, Hong Kong SAR, China. (2)Department of Electrical

Engineering, Princeton University, Princeton, New Jersey, United States of

America.

Protein subcellular localization prediction, as an essential step to elucidate

the functions in vivo of proteins and identify drugs targets, has been

extensively studied in previous decades. Instead of only determining subcellular

localization of single-label proteins, recent studies have focused on predicting

both single- and multi-location proteins. Computational methods based on Gene

Ontology (GO) have been demonstrated to be superior to methods based on other

features. However, existing GO-based methods focus on the occurrences of GO terms

and disregard their relationships. This paper proposes a multi-label

subcellular-localization predictor, namely HybridGO-Loc, that leverages not only

the GO term occurrences but also the inter-term relationships. This is achieved

by hybridizing the GO frequencies of occurrences and the semantic similarity

between GO terms. Given a protein, a set of GO terms are retrieved by searching

against the gene ontology database, using the accession numbers of homologous

proteins obtained via BLAST search as the keys. The frequency of GO occurrences

and semantic similarity (SS) between GO terms are used to formulate frequency

vectors and semantic similarity vectors, respectively, which are subsequently

hybridized to construct fusion vectors. An adaptive-decision based multi-label

support vector machine (SVM) classifier is proposed to classify the fusion

vectors. Experimental results based on recent benchmark datasets and a new

dataset containing novel proteins show that the proposed hybrid-feature predictor

significantly outperforms predictors based on individual GO features as well as

other state-of-the-art predictors. For readers' convenience, the HybridGO-Loc

server, which is for predicting virus or plant proteins, is available online at

http://bioinfo.eie.polyu.edu.hk/HybridGoServer/.

DOI: 10.1371/journal.pone.0089545

PMCID: PMC3960097

PMID: 24647341 [Indexed for MEDLINE]

876. Database (Oxford). 2014 Mar 18;2014:bau020. doi: 10.1093/database/bau020. Print

2014.

Comparison of sequence variants in transcriptomic control regions across 17 mouse

genomes.

Nguyen C(1), Baten A, Morahan G.

Author information:

(1)Centre for Diabetes Research, The Western Australian Institute for Medical

Research, Western Australia, Australia, Centre of Medical Research, University of

Western Australia, Perth, Western Australia, Australia and Southern Cross Plant

Science, Southern Cross University, Lismore, New South Wales, Australia.

The laboratory mouse is the most widely used mammalian model organism in

biomedical research, so a thorough annotation of functional variation in the

mouse genome would be of significant value. In this study, we compared sequence

variation in a comprehensive list of functional elements (e.g. promoters,

enhancers and CTCF binding sites) across 17 inbred mouse strains. Sequences were

derived for ∼300,000 functional elements experimentally identified by the mouse

ENCODE project as regulating gene expression in 19 different tissue sources. We

aligned sequences for each predicted cis-regulatory element to genomes of 17

mouse strains. This yielded a database comprising ∼5 million aligned sequences,

allowing interrogation of sequence variation of functional elements for each of

the 19 tissues/cell types in commonly used mouse strains. We also developed an

online tool to visualize the genome around each predicted cis-regulatory element

in each tissue context and which allows efficient comparison of variation between

any two sets of strains. This will be particularly useful in the context of the

Collaborative Cross (CC), which was conceived as a powerful new systems genetics

resource to accelerate gene discovery. Comprising a large number of inbred

strains derived from eight genetically diverse founders, the CC offers rapid

mapping and identification of genes that mediate complex traits. We show that,

among the 17 sequenced strains, the set of CC founder strains captures the most

variability in the ENCODE elements, further emphasizing the value of this

resource. Database URL: www.sysgen.org/ecco.

DOI: 10.1093/database/bau020

PMCID: PMC3958616

PMID: 24647628 [Indexed for MEDLINE]

877. J Med Internet Res. 2014 Mar 18;16(3):e86. doi: 10.2196/jmir.3069.

Do email and mobile phone prompts stimulate primary school children to reuse an

Internet-delivered smoking prevention intervention?

Cremers HP(1), Mercken L, Crutzen R, Willems P, de Vries H, Oenema A.

Author information:

(1)School for Public Health and Primary Care (CAPHRI), Department of Health

Promotion, Maastricht University, Maastricht, Netherlands.

p.cremers@maastrichtuniversity.nl.

BACKGROUND: Improving the use (eg, initial visit and revisits) of

Internet-delivered interventions to promote healthy lifestyles such as

non-smoking is one of the largest challenges in the field of eHealth. Prompts

have shown to be effective in stimulating reuse of Internet-delivered

interventions among adults and adolescents. However, evidence concerning

effectiveness of prompts to promote reuse of a website among children is still

scarce.

OBJECTIVE: The aim of this study is to investigate (1) whether prompts are

effective in promoting reuse of an intervention website containing information on

smoking prevention for children, (2) whether the content of the prompt is

associated with its effect in terms of reuse, and (3) whether there are

differences between children who do or do not respond to prompts.

METHODS: The sample of this cluster-randomized study consisted of 1124 children

(aged 10-11 years) from 108 Dutch primary schools, who were assigned to the

experimental group of an Internet-delivered smoking prevention intervention

study. All participants completed a Web-based questionnaire on factors related to

(non-)smoking. Schools were randomized to a no-prompt group (n=50) or a prompt

group (n=58). All children could revisit the intervention website, but only the

children in the prompt group received email and SMS prompts to revisit the

website. Those prompt messages functioned as a teaser to stimulate reuse of the

intervention website. Reuse of the website was objectively tracked by means of a

server registration system. Repeated measures analysis of variance and linear

regression analysis were performed to assess the effects of prompts on website

reuse and to identify individual characteristics of participants who reuse the

intervention website.

RESULTS: Children in the prompt group reused the intervention website

significantly more often compared to children in the no-prompt group (B=1.56,

P<.001). Prompts announcing new animated videos (F1,1122=9.33, P=.002) and games

about (non-)smoking on the website (F1,1122=8.28, P=.004) resulted in most reuse

of the website. Within the prompt group, children with a low socioeconomic status

(SES) reused the intervention website more often (B=2.19, P<.001) than children

of high SES (B=0.93, P=.005).

CONCLUSIONS: Prompts can stimulate children to reuse an intervention website

aimed at smoking prevention. Prompts showed, furthermore, to stimulate children

of a low SES slightly more to reuse an intervention website, which is often a

difficult target group in terms of stimulating participation. However, the number

of revisits was quite low, which requires further study into how prompts can be

optimized in terms of content and frequency to improve the number of revisits.

TRIAL REGISTRATION: Netherlands Trial Register Number: NTR3116;

http://www.trialregister.nl/trialreg/admin/rctview.asp?TC=3116 (Archived by

WebCite at http://www.webcitation.org/6O0wQYuPI).

DOI: 10.2196/jmir.3069

PMCID: PMC3978553

PMID: 24642082 [Indexed for MEDLINE]

878. Anal Biochem. 2014 Mar 15;449:164-71. doi: 10.1016/j.ab.2013.12.013. Epub 2013

Dec 21.

MSLoc-DT: a new method for predicting the protein subcellular location of

multispecies based on decision templates.

Zhang SW(1), Liu YF(2), Yu Y(2), Zhang TH(2), Fan XN(2).

Author information:

(1)College of Automation, Northwestern Polytechnical University, Xi'an 710072,

China. Electronic address: zhangsw@nwpu.edu.cn. (2)College of Automation,

Northwestern Polytechnical University, Xi'an 710072, China.

Revealing the subcellular location of newly discovered protein sequences can

bring insight to their function and guide research at the cellular level. The

rapidly increasing number of sequences entering the genome databanks has called

for the development of automated analysis methods. Currently, most existing

methods used to predict protein subcellular locations cover only one, or a very

limited number of species. Therefore, it is necessary to develop reliable and

effective computational approaches to further improve the performance of protein

subcellular prediction and, at the same time, cover more species. The current

study reports the development of a novel predictor called MSLoc-DT to predict the

protein subcellular locations of human, animal, plant, bacteria, virus, fungi,

and archaea by introducing a novel feature extraction approach termed Amino Acid

Index Distribution (AAID) and then fusing gene ontology information, sequential

evolutionary information, and sequence statistical information through four

different modes of pseudo amino acid composition (PseAAC) with a decision

template rule. Using the jackknife test, MSLoc-DT can achieve 86.5, 98.3, 90.3,

98.5, 95.9, 98.1, and 99.3% overall accuracy for human, animal, plant, bacteria,

virus, fungi, and archaea, respectively, on seven stringent benchmark datasets.

Compared with other predictors (e.g., Gpos-PLoc, Gneg-PLoc, Virus-PLoc,

Plant-PLoc, Plant-mPLoc, ProLoc-Go, Hum-PLoc, GOASVM) on the gram-positive,

gram-negative, virus, plant, eukaryotic, and human datasets, the new MSLoc-DT

predictor is much more effective and robust. Although the MSLoc-DT predictor is

designed to predict the single location of proteins, our method can be extended

to multiple locations of proteins by introducing multilabel machine learning

approaches, such as the support vector machine and deep learning, as substitutes

for the K-nearest neighbor (KNN) method. As a user-friendly web server, MSLoc-DT

is freely accessible at http://bioinfo.ibp.ac.cn/MSLOC\_DT/index.html.

Crown Copyright © 2013. Published by Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2013.12.013

PMID: 24361712 [Indexed for MEDLINE]

879. Bioinformatics. 2014 Mar 15;30(6):889-90. doi: 10.1093/bioinformatics/btt645.

Epub 2013 Nov 7.

PREDDIMER: a web server for prediction of transmembrane helical dimers.

Polyansky AA(1), Chugunov AO, Volynsky PE, Krylov NA, Nolde DE, Efremov RG.

Author information:

(1)M.M. Shemyakin and Yu.A. Ovchinnikov Institute of Bioorganic Chemistry,

Russian Academy of Sciences, Moscow 117997, Russia, Max F. Perutz Laboratories,

Department of Structural and Computational Biology, University of Vienna, Campus

Vienna Biocenter 5, Vienna AT-1030, Austria and Moscow Institute of Physics and

Technology (State University), Dolgoprudny, Moscow Region 141700, Russia.

SUMMARY: Here we present PREDDIMER, a web tool for prediction of dimer structure

of transmembrane (TM) helices. PREDDIMER allows (i) reconstruction of a number of

dimer structures for given sequence(s) of TM protein fragments, (ii) ranking and

filtering of predicted structures according to respective values of a scoring

function, (iii) visualization of predicted 3D dimer structures and (iv)

visualization of surface hydrophobicity of TM helices and their contacting

(interface) regions represented as 2D maps.

RESULTS: We implemented online the original PREDDIMER algorithm and benchmarked

the server on 11 TM sequences, whose 3D dimer conformations were obtained

previously by nuclear magnetic resonance spectroscopy. In the most of tested

cases backbone root-mean-square deviations of closest predicted conformations

from the experimental reference are below 3 Å. A randomization test displays good

anticorrelation (-0.82) between values of the scoring function and statistical

significance of the prediction 'by chance'. Going beyond a single dimer

conformation, our web tool predicts an ensemble of possible conformations, which

may be useful for explanation of a functioning of bitopic membrane proteins, e.g.

receptor tyrosine kinases.

AVAILABILITY AND IMPLEMENTATION: PREDDIMER can be accessed for free on the web at

http://model.nmr.ru/preddimer/

CONTACT: newant@gmail.com

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt645

PMID: 24202542 [Indexed for MEDLINE]

880. CPT Pharmacometrics Syst Pharmacol. 2014 Mar 12;3:e105. doi: 10.1038/psp.2014.1.

VNP: Interactive Visual Network Pharmacology of Diseases, Targets, and Drugs.

Hu QN(1), Deng Z(1), Tu W(1), Yang X(1), Meng ZB(2), Deng ZX(1), Liu J(2).

Author information:

(1)Key Laboratory of Combinatorial Biosynthesis and Drug Discovery (Wuhan

University), Ministry of Education, and Wuhan University School of Pharmaceutical

Sciences, Wuhan, P. R. China. (2)State Key Laboratory of Software Engineering

(Wuhan University) and Wuhan University School of Computer Science, Wuhan, P. R.

China.

In drug discovery, promiscuous targets, multifactorial diseases, and "dirty"

drugs construct complex network relationships. Network pharmacology description

and analysis not only give a systems-level understanding of drug action and

disease complexity but can also help to improve the efficiency of target

selection and drug design. Visual network pharmacology (VNP) is developed to

visualize network pharmacology of targets, diseases, and drugs with a graph

network by using disease, target or drug names, chemical structures, or protein

sequence. To our knowledge, VNP is the first free interactive VNP server that

should be very helpful for systems pharmacology research. VNP is freely available

at http://cadd.whu.edu.cn/ditad/vnpsearch.

DOI: 10.1038/psp.2014.1

PMCID: PMC4039393

PMID: 24622768

881. J Theor Biol. 2014 Mar 7;344:31-9. doi: 10.1016/j.jtbi.2013.11.017. Epub 2013 Dec

4.

A two-stage SVM method to predict membrane protein types by incorporating amino

acid classifications and physicochemical properties into a general form of Chou's

PseAAC.

Han GS(1), Yu ZG(2), Anh V(3).

Author information:

(1)School of Mathematics and Computational Science, Xiangtan University, Hunan

411105, China. (2)School of Mathematics and Computational Science, Xiangtan

University, Hunan 411105, China; School of Mathematical Science, Queensland

University of Technology, GPO Box 2434, Brisbane Q 4001, Australia. Electronic

address: yuzg1970@yahoo.com. (3)School of Mathematical Science, Queensland

University of Technology, GPO Box 2434, Brisbane Q 4001, Australia.

Membrane proteins play important roles in many biochemical processes and are also

attractive targets of drug discovery for various diseases. The elucidation of

membrane protein types provides clues for understanding the structure and

function of proteins. Recently we developed a novel system for predicting protein

subnuclear localizations. In this paper, we propose a simplified version of our

system for predicting membrane protein types directly from primary protein

structures, which incorporates amino acid classifications and physicochemical

properties into a general form of pseudo-amino acid composition. In this

simplified system, we will design a two-stage multi-class support vector machine

combined with a two-step optimal feature selection process, which proves very

effective in our experiments. The performance of the present method is evaluated

on two benchmark datasets consisting of five types of membrane proteins. The

overall accuracies of prediction for five types are 93.25% and 96.61% via the

jackknife test and independent dataset test, respectively. These results indicate

that our method is effective and valuable for predicting membrane protein types.

A web server for the proposed method is available at

http://www.juemengt.com/jcc/memty\_page.php.

Copyright © 2013 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2013.11.017

PMID: 24316387 [Indexed for MEDLINE]

882. Algorithms Mol Biol. 2014 Mar 6;9(1):5. doi: 10.1186/1748-7188-9-5.

Faster algorithms for RNA-folding using the Four-Russians method.

Venkatachalam B(1), Gusfield D, Frid Y.

Author information:

(1)Department of Computer Science, University of California, Davis, 1 Shields

Ave, Davis, CA, USA. balaji@cs.ucdavis.edu.

BACKGROUND: The secondary structure that maximizes the number of non-crossing

matchings between complimentary bases of an RNA sequence of length n can be

computed in O(n3) time using Nussinov's dynamic programming algorithm. The

Four-Russians method is a technique that reduces the running time for certain

dynamic programming algorithms by a multiplicative factor after a preprocessing

step where solutions to all smaller subproblems of a fixed size are exhaustively

enumerated and solved. Frid and Gusfield designed an O(n3logn) algorithm for RNA

folding using the Four-Russians technique. In their algorithm the preprocessing

is interleaved with the algorithm computation.

THEORETICAL RESULTS: We simplify the algorithm and the analysis by doing the

preprocessing once prior to the algorithm computation. We call this the

two-vector method. We also show variants where instead of exhaustive

preprocessing, we only solve the subproblems encountered in the main algorithm

once and memoize the results. We give a simple proof of correctness and explore

the practical advantages over the earlier method.The Nussinov algorithm admits an

O(n2) time parallel algorithm. We show a parallel algorithm using the two-vector

idea that improves the time bound to O(n2logn).

PRACTICAL RESULTS: We have implemented the parallel algorithm on graphics

processing units using the CUDA platform. We discuss the organization of the data

structures to exploit coalesced memory access for fast running times. The ideas

to organize the data structures also help in improving the running time of the

serial algorithms. For sequences of length up to 6000 bases the parallel

algorithm takes only about 2.5 seconds and the two-vector serial method takes

about 57 seconds on a desktop and 15 seconds on a server. Among the serial

algorithms, the two-vector and memoized versions are faster than the

Frid-Gusfield algorithm by a factor of 3, and are faster than Nussinov by up to a

factor of 20. The source-code for the algorithms is available at

http://github.com/ijalabv/FourRussiansRNAFolding.

DOI: 10.1186/1748-7188-9-5

PMCID: PMC3996002

PMID: 24602450

883. Anal Chem. 2014 Mar 4;86(5):2510-20. doi: 10.1021/ac403544k. Epub 2014 Feb 12.

Quantitative structure-property relationship modeling: a valuable support in

high-throughput screening quality control.

Ruggiu F(1), Gizzi P, Galzi JL, Hibert M, Haiech J, Baskin I, Horvath D, Marcou

G, Varnek A.

Author information:

(1)Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg , 1

rue Blaise Pascal, 67000 Strasbourg, France.

Evaluation of important pharmacokinetic properties such as hydrophobicity by

high-throughput screening (HTS) methods is a major issue in drug discovery. In

this paper, we present measurements of the chromatographic hydrophobicity index

(CHI) on a subset of the French chemical library Chimiothèque Nationale (CN). The

data were used in quantitative structure-property relationship (QSPR) modeling in

order to annotate the CN. An algorithm is proposed to detect problematic

molecules with large prediction errors, called outliers. In order to find an

explanation for these large discrepancies between predicted and experimental

values, these compounds were reanalyzed experimentally. As the first selected

outliers indeed had experimental problems, including hydrolysis or sheer absence

of expected structure, we herewith propose the use of QSPR as a support tool for

quality control of screening data and encourage cooperation between experimental

and theoretical teams to improve results. The corrected data were used to produce

a model, which is freely available on our web server at

http://infochim.u-strasbg.fr/webserv/VSEngine.html .

DOI: 10.1021/ac403544k

PMID: 24479843 [Indexed for MEDLINE]

884. Acta Crystallogr A Found Adv. 2014 Mar;70(Pt 2):126-37. doi:

10.1107/S205327331303091X. Epub 2014 Feb 12.

Brillouin-zone database on the Bilbao Crystallographic Server.

Aroyo MI(1), Orobengoa D(1), de la Flor G(1), Tasci ES(2), Perez-Mato JM(1),

Wondratschek H(3).

Author information:

(1)Departamento Física de la Materia Condensada, Universidad del País Vasco

(UPV/EHU), Bilbao, Spain. (2)Physics Department, Middle East Technical

University, Ankara, Turkey. (3)Laboratorium für Applikationen der

Synchrotronstrahlung (LAS), Universität Karlsruhe, Germany.

The Brillouin-zone database of the Bilbao Crystallographic Server

(http://www.cryst.ehu.es) offers k-vector tables and figures which form the

background of a classification of the irreducible representations of all 230

space groups. The symmetry properties of the wavevectors are described by the

so-called reciprocal-space groups and this classification scheme is compared with

the classification of Cracknell et al. [Kronecker Product Tables, Vol. 1, General

Introduction and Tables of Irreducible Representations of Space Groups (1979).

New York: IFI/Plenum]. The compilation provides a solution to the problems of

uniqueness and completeness of space-group representations by specifying the

independent parameter ranges of general and special k vectors. Guides to the

k-vector tables and figures explain the content and arrangement of the data.

Recent improvements and modifications of the Brillouin-zone database, including

new tables and figures for the trigonal, hexagonal and monoclinic space groups,

are discussed in detail and illustrated by several examples.

DOI: 10.1107/S205327331303091X

PMID: 24572313

885. Acta Pharmacol Sin. 2014 Mar;35(3):419-31. doi: 10.1038/aps.2013.153. Epub 2014

Feb 3.

Multi-algorithm and multi-model based drug target prediction and web server.

Liu YT(1), Li Y(1), Huang ZF(1), Xu ZJ(1), Yang Z(1), Chen ZX(1), Chen KX(1), Shi

JY(2), Zhu WL(1).

Author information:

(1)Drug Discovery and Design Center, Key Laboratory of Receptor Research,

Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai

201203, China. (2)Informatics Department, UCB Pharma, 216 Bath Road, Slough SL1

4EN, UK.

AIM: To develop a reliable computational approach for predicting potential drug

targets based merely on protein sequence.

METHODS: With drug target and non-target datasets prepared and 3 classification

algorithms (Support Vector Machine, Neural Network and Decision Tree), a

multi-algorithm and multi-model based strategy was employed for constructing

models to predict potential drug targets.

RESULTS: Twenty one prediction models for each of the 3 algorithms were

successfully developed. Our evaluation results showed that ∼30% of human proteins

were potential drug targets, and ∼40% of putative targets for the drugs

undergoing phase II clinical trials were probably non-targets. A public web

server named D3TPredictor (http://www.d3pharma.com/d3tpredictor) was constructed

to provide easy access.

CONCLUSION: Reliable and robust drug target prediction based on protein sequences

is achieved using the multi-algorithm and multi-model strategy.

DOI: 10.1038/aps.2013.153

PMCID: PMC4647888

PMID: 24487966 [Indexed for MEDLINE]

886. Behav Res Methods. 2014 Mar;46(1):148-58. doi: 10.3758/s13428-013-0356-8.

Phi-square Lexical Competition Database (Phi-Lex): an online tool for quantifying

auditory and visual lexical competition.

Strand JF(1).

Author information:

(1)Department of Psychology, Carleton College, Northfield, MN, 55057, USA,

jstrand@carleton.edu.

A widely agreed-upon feature of spoken word recognition is that multiple lexical

candidates in memory are simultaneously activated in parallel when a listener

hears a word, and that those candidates compete for recognition (Luce, Goldinger,

Auer, & Vitevitch, Perception 62:615-625, 2000; Luce & Pisoni, Ear and Hearing

19:1-36, 1998; McClelland & Elman, Cognitive Psychology 18:1-86, 1986). Because

the presence of those competitors influences word recognition, much research has

sought to quantify the processes of lexical competition. Metrics that quantify

lexical competition continuously are more effective predictors of auditory and

visual (lipread) spoken word recognition than are the categorical metrics

traditionally used (Feld & Sommers, Speech Communication 53:220-228, 2011; Strand

& Sommers, Journal of the Acoustical Society of America 130:1663-1672, 2011). A

limitation of the continuous metrics is that they are somewhat computationally

cumbersome and require access to existing speech databases. This article

describes the Phi-square Lexical Competition Database (Phi-Lex): an online,

searchable database that provides access to multiple metrics of auditory and

visual (lipread) lexical competition for English words, available at

www.juliastrand.com/phi-lex .

DOI: 10.3758/s13428-013-0356-8

PMID: 23754576 [Indexed for MEDLINE]

887. J Virol Methods. 2014 Mar;198:41-55. doi: 10.1016/j.jviromet.2013.12.012. Epub

2013 Dec 31.

WNV Typer: a server for genotyping of West Nile viruses using an alignment-free

method based on a return time distribution.

Kolekar P(1), Hake N(2), Kale M(3), Kulkarni-Kale U(4).

Author information:

(1)Bioinformatics Centre, University of Pune, Pune 411007, India. Electronic

address: pandurang@bioinfo.net.in. (2)Bioinformatics Centre, University of Pune,

Pune 411007, India. (3)Department of Statistics, University of Pune, Pune 411007,

India. Electronic address: mmkale@stats.unipune.ac.in. (4)Bioinformatics Centre,

University of Pune, Pune 411007, India. Electronic address:

urmila@bioinfo.net.in.

West Nile virus (WNV), genus Flavivirus, family Flaviviridae, is a major cause of

viral encephalitis with broad host range and global spread. The virus has

undergone a series of evolutionary changes with emergence of various genotypic

lineages that are known to differ in type and severity of the diseases caused.

Currently, genotyping is carried out using molecular phylogeny of complete coding

sequences and genotype is assigned based on proximity to reference genotypes in

tree topology. Efficient epidemiological surveillance of WNVs demands development

of objective criteria for typing. An alignment-free approach based on return time

distribution (RTD) of k-mers has been validated for genotyping of WNVs. The RTDs

of complete genome sequences at k=7 were found to be optimum for classification

of the known lineages of WNVs as well as for genotyping. It provides time and

computationally efficient alternative for genome based annotation of WNV

lineages. The development of a WNV Typer server based on RTD is described

(http://bioinfo.net.in/wnv/homepage.html). Both the method and the server have

100% sensitivity and specificity.

Copyright © 2014 The Authors. Published by Elsevier B.V. All rights reserved.

DOI: 10.1016/j.jviromet.2013.12.012

PMID: 24388930 [Indexed for MEDLINE]

888. Mol Inform. 2014 Mar;33(3):230-9. doi: 10.1002/minf.201300077. Epub 2014 Mar 11.

Prediction of Signal Peptide Cleavage Sites with Subsite-Coupled and Template

Matching Fusion Algorithm.

Zhang SW(1,)(2), Zhang TH(3), Zhang JN(3), Huang Y(4).

Author information:

(1)School of Automation, Northwestern Polytechnical University, Xi'an, 710072,

P. R. China. zhangsw@nwpu.edu.cn. (2)Key Laboratory of Information Fusion

Technology, Ministry of Education, 710072, Xi'an, P. R. China.

zhangsw@nwpu.edu.cn. (3)School of Automation, Northwestern Polytechnical

University, Xi'an, 710072, P. R. China. (4)Department of ECE, University of Texas

at San Antonio, San Antonio, TX, USA. yufei.huang@utsa.edu.

Fast and effective prediction of signal peptides (SP) and their cleavage sites is

of great importance in computational biology. The approaches developed to predict

signal peptide can be roughly divided into machine learning based, and sliding

windows based. In order to further increase the prediction accuracy and coverage

of organism for SP cleavage sites, we propose a novel method for predicting SP

cleavage sites called Signal-CTF that utilizes machine learning and sliding

windows, and is designed for N-termial secretory proteins in a large variety of

organisms including human, animal, plant, virus, bacteria, fungi and archaea.

Signal-CTF consists of three distinct elements: (1) a subsite-coupled and

regularization function with a scaled window of fixed width that selects a set of

candidates of possible secretion-cleavable segment for a query secretory protein;

(2) a sum fusion system that integrates the outcomes from aligning the cleavage

site template sequence with each of the aforementioned candidates in a scaled

window of fixed width to determine the best candidate cleavage sites for the

query secretory protein; (3) a voting system that identifies the ultimate signal

peptide cleavage site among all possible results derived from using scaled

windows of different width. When compared with Signal-3L and SignalP 4.0

predictors, the prediction accuracy of Signal-CTF is 4-12 %, 10-25 % higher than

that of Signal-3L for human, animal and eukaryote, and SignalP 4.0 for eukaryota,

Gram-positive bacteria and Gram-negative bacteria, respectively. Comparing with

PRED-SIGNAL and SignalP 4.0 predictors on the 32 archaea secretory proteins of

used in Bagos's paper, the prediction accuracy of Signal-CTF is 12.5 %, 25 %

higher than that of PRED-SIGNAL and SignalP 4.0, respectively. The predicting

results of several long signal peptides show that the Signal-CTF can better

predict cleavage sites for long signal peptides than SignalP, Phobius, Philius,

SPOCTOPUS, Signal-CF and Signal-3L. These results show that Signal-CTF is more

accurate and flexible in predicting signal peptides of different characteristics

for many organisms. Signal-CTF is freely available as a web-server at

http://darwin2.cbi.utsa.edu/minniweb/index.html.

© 2014 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

DOI: 10.1002/minf.201300077

PMID: 27485691

889. BMC Bioinformatics. 2014 Feb 28;15:61. doi: 10.1186/1471-2105-15-61.

CDSbank: taxonomy-aware extraction, selection, renaming and formatting of

protein-coding DNA or amino acid sequences.

Hazes B(1).

Author information:

(1)Department of Medical Microbiology & Immunology, 6-020 Katz Group Centre,

University of Alberta, Edmonton, Alberta T6G 2E1, Canada. bart.hazes@ualberta.ca.

BACKGROUND: Protein-coding DNA sequences and their corresponding amino acid

sequences are routinely used to study relationships between sequence, structure,

function, and evolution. The rapidly growing size of sequence databases increases

the power of such comparative analyses but it makes it more challenging to

prepare high quality sequence data sets with control over redundancy, quality,

completeness, formatting, and labeling. Software tools for some individual steps

in this process exist but manual intervention remains a common and time consuming

necessity.

DESCRIPTION: CDSbank is a database that stores both the protein-coding DNA

sequence (CDS) and amino acid sequence for each protein annotated in Genbank.

CDSbank also stores Genbank feature annotation, a flag to indicate incomplete 5'

and 3' ends, full taxonomic data, and a heuristic to rank the scientific interest

of each species. This rich information allows fully automated data set

preparation with a level of sophistication that aims to meet or exceed manual

processing. Defaults ensure ease of use for typical scenarios while allowing

great flexibility when needed. Access is via a free web server at

http://hazeslab.med.ualberta.ca/CDSbank/.

CONCLUSIONS: CDSbank presents a user-friendly web server to download, filter,

format, and name large sequence data sets. Common usage scenarios can be accessed

via pre-programmed default choices, while optional sections give full control

over the processing pipeline. Particular strengths are: extract protein-coding

DNA sequences just as easily as amino acid sequences, full access to taxonomy for

labeling and filtering, awareness of incomplete sequences, and the ability to

take one protein sequence and extract all synonymous CDS or identical protein

sequences in other species. Finally, CDSbank can also create labeled property

files to, for instance, annotate or re-label phylogenetic trees.

DOI: 10.1186/1471-2105-15-61

PMCID: PMC3942066

PMID: 24580755 [Indexed for MEDLINE]

890. Parasit Vectors. 2014 Feb 21;7:76. doi: 10.1186/1756-3305-7-76.

An online tool for mapping insecticide resistance in major Anopheles vectors of

human malaria parasites and review of resistance status for the Afrotropical

region.

Knox TB, Juma EO, Ochomo EO, Pates Jamet H, Ndungo L, Chege P, Bayoh NM,

N'Guessan R, Christian RN, Hunt RH, Coetzee M(1).

Author information:

(1)Wits Research Institute for Malaria, School of Pathology, Faculty of Health

Sciences, University of the Witwatersrand, Johannesburg, South Africa.

maureenc@nicd.ac.za.

BACKGROUND: Malaria control programmes across Africa and beyond are facing

increasing insecticide resistance in the major anopheline vectors. In order to

preserve or prolong the effectiveness of the main malaria vector interventions,

up-to-date and easily accessible insecticide resistance data that are

interpretable at operationally-relevant scales are critical. Herein we introduce

and demonstrate the usefulness of an online mapping tool, IR Mapper.

METHODS: A systematic search of published, peer-reviewed literature was performed

and Anopheles insecticide susceptibility and resistance mechanisms data were

extracted and added to a database after a two-level verification process. IR

Mapper ( http://www.irmapper.com) was developed using the ArcGIS for JavaScript

Application Programming Interface and ArcGIS Online platform for exploration and

projection of these data.

RESULTS: Literature searches yielded a total of 4,084 susceptibility data points

for 1,505 populations, and 2,097 resistance mechanisms data points for 1,000

populations of Anopheles spp. tested via recommended WHO methods from 54

countries between 1954 and 2012. For the Afrotropical region, data were most

abundant for populations of An. gambiae, and pyrethroids and DDT were more often

used in susceptibility assays (51.1 and 26.8% of all reports, respectively) than

carbamates and organophosphates. Between 2001 and 2012, there was a clear

increase in prevalence and distribution of confirmed resistance of An. gambiae

s.l. to pyrethroids (from 41 to 87% of the mosquito populations tested) and DDT

(from 64 to 91%) throughout the Afrotropical region. Metabolic resistance

mechanisms were detected in western and eastern African populations and the two

kdr mutations (L1014S and L1014F) were widespread. For An. funestus s.l.,

relatively few populations were tested, although in 2010-2012 resistance was

reported in 50% of 10 populations tested. Maps are provided to illustrate the use

of IR Mapper and the distribution of insecticide resistance in malaria vectors in

Africa.

CONCLUSIONS: The increasing pyrethroid and DDT resistance in Anopheles in the

Afrotropical region is alarming. Urgent attention should be afforded to testing

An. funestus populations especially for metabolic resistance mechanisms. IR

Mapper is a useful tool for investigating temporal and spatial trends in

Anopheles resistance to support the pragmatic use of insecticidal interventions.

DOI: 10.1186/1756-3305-7-76

PMCID: PMC3942210

PMID: 24559061 [Indexed for MEDLINE]

891. PLoS One. 2014 Feb 21;9(2):e85412. doi: 10.1371/journal.pone.0085412. eCollection

2014.

Energy parameters and novel algorithms for an extended nearest neighbor energy

model of RNA.

Dotu I(1), Mechery V(2), Clote P(1).

Author information:

(1)Biology Department, Boston College, Chestnut Hill, Massachusetts, United

States of America. (2)Hofstra North Shore-LIJ School of Medicine, Hempstead, New

York, United States of America.

We describe the first algorithm and software, RNAenn, to compute the partition

function and minimum free energy secondary structure for RNA with respect to an

extended nearest neighbor energy model. Our next-nearest-neighbor triplet energy

model appears to lead to somewhat more cooperative folding than does the nearest

neighbor energy model, as judged by melting curves computed with RNAenn and with

two popular software implementations for the nearest-neighbor energy model. A web

server is available at http://bioinformatics.bc.edu/clotelab/RNAenn/.

DOI: 10.1371/journal.pone.0085412

PMCID: PMC3931620

PMID: 24586240 [Indexed for MEDLINE]

892. PLoS One. 2014 Feb 20;9(2):e89575. doi: 10.1371/journal.pone.0089575. eCollection

2014.

LAceP: lysine acetylation site prediction using logistic regression classifiers.

Hou T(1), Zheng G(2), Zhang P(3), Jia J(4), Li J(4), Xie L(4), Wei C(5), Li Y(1).

Author information:

(1)School of Biological Engineering, East China University of Science and

Technology, Shanghai, China ; Shanghai Center for Bioinformation Technology,

Shanghai, China ; Key Laboratory of Systems Biology, Shanghai Institutes for

Biological Sciences, Chinese Academy of Sciences, Shanghai, China. (2)CAS-MPG

Partner Institute for Computational Biology, Shanghai Institutes for Biological

Sciences, Chinese Academy of Sciences, Shanghai, China. (3)Key Laboratory of

Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of

Sciences, Shanghai, China. (4)Shanghai Center for Bioinformation Technology,

Shanghai, China. (5)Shanghai Center for Bioinformation Technology, Shanghai,

China ; School of Life Sciences and Biotechnology, Shanghai Jiao Tong University,

Shanghai, China.

BACKGROUND: Lysine acetylation is a crucial type of protein post-translational

modification, which is involved in many important cellular processes and serious

diseases. However, identification of protein acetylated sites through traditional

experiment methods is time-consuming and laborious. Those methods are not

suitable to identify a large number of acetylated sites quickly. Therefore,

computational methods are still very valuable to accelerate lysine acetylated

site finding.

RESULT: In this study, many biological characteristics of acetylated sites have

been investigated, such as the amino acid sequence around the acetylated sites,

the physicochemical property of the amino acids and the transition probability of

adjacent amino acids. A logistic regression method was then utilized to integrate

these information for generating a novel lysine acetylation prediction system

named LAceP. When compared with existing methods, LAceP overwhelms most of

state-of-the-art methods. Especially, LAceP has a more balanced prediction

capability for positive and negative datasets.

CONCLUSION: LAceP can integrate different biological features to predict lysine

acetylation with high accuracy. An online web server is freely available at

http://www.scbit.org/iPTM/.

DOI: 10.1371/journal.pone.0089575

PMCID: PMC3930742

PMID: 24586884 [Indexed for MEDLINE]

893. Bioinformation. 2014 Feb 19;10(2):98-100. doi: 10.6026/97320630010098.

eCollection 2014.

Insect barcode information system.

Pratheepa M(1), Jalali SK(1), Arokiaraj RS(1), Venkatesan T(1), Nagesh M(1),

Panda M(1), Pattar S(1).

Author information:

(1)National Bureau of Agriculturally Important Insects, Post Bag No. 2491, H.A.

Farm Post, Hebbal, Bellary Road, Bangalore - 560 024, Karnataka, India.

Insect Barcode Information System called as Insect Barcode Informática (IBIn) is

an online database resource developed by the National Bureau of Agriculturally

Important Insects, Bangalore. This database provides acquisition, storage,

analysis and publication of DNA barcode records of agriculturally important

insects, for researchers specifically in India and other countries. It bridges a

gap in bioinformatics by integrating molecular, morphological and distribution

details of agriculturally important insects. IBIn was developed using PHP/My SQL

by using relational database management concept. This database is based on the

client- server architecture, where many clients can access data simultaneously.

IBIn is freely available on-line and is user-friendly. IBIn allows the registered

users to input new information, search and view information related to DNA

barcode of agriculturally important insects.This paper provides a current status

of insect barcode in India and brief introduction about the database

IBIn.AVAILABILITY: http://www.nabg-nbaii.res.in/barcode.

DOI: 10.6026/97320630010098

PMCID: PMC3937583

PMID: 24616562

894. PLoS One. 2014 Feb 19;9(2):e89246. doi: 10.1371/journal.pone.0089246. eCollection

2014.

PalmPred: an SVM based palmitoylation prediction method using sequence profile

information.

Kumari B(1), Kumar R(1), Kumar M(1).

Author information:

(1)Department of Biophysics, University of Delhi South Campus, New Delhi, India.

Protein palmitoylation is the covalent attachment of the 16-carbon fatty acid

palmitate to a cysteine residue. It is the most common acylation of protein and

occurs only in eukaryotes. Palmitoylation plays an important role in the

regulation of protein subcellular localization, stability, translocation to lipid

rafts and many other protein functions. Hence, the accurate prediction of

palmitoylation site(s) can help in understanding the molecular mechanism of

palmitoylation and also in designing various related experiments. Here we present

a novel in silico predictor called 'PalmPred' to identify palmitoylation sites

from protein sequence information using a support vector machine model. The best

performance of PalmPred was obtained by incorporating sequence conservation

features of peptide of window size 11 using a leave-one-out approach. It helped

in achieving an accuracy of 91.98%, sensitivity of 79.23%, specificity of 94.30%,

and Matthews Correlation Coefficient of 0.71. PalmPred outperformed existing

palmitoylation site prediction methods - IFS-Palm and WAP-Palm on an independent

dataset. Based on these measures it can be anticipated that PalmPred will be

helpful in identifying candidate palmitoylation sites. All the source datasets,

standalone and web-server are available at http://14.139.227.92/mkumar/palmpred/.

DOI: 10.1371/journal.pone.0089246

PMCID: PMC3929663

PMID: 24586628 [Indexed for MEDLINE]

895. Database (Oxford). 2014 Feb 17;2014:bau005. doi: 10.1093/database/bau005. Print

2014.

MICdb3.0: a comprehensive resource of microsatellite repeats from prokaryotic

genomes.

Mudunuri SB(1), Patnana S, Nagarajaram HA.

Author information:

(1)Department of Computer Science & Engineering, Grandhi Varalakshmi Venkatarao

Institute of Technology, Bhimavaram, Andhra Pradesh 534 207, India, Training &

Delivery Department, TalentSprint Educational Services, IIIT Campus, Hyderabad,

Andhra Pradesh 500 032, India and Laboratory of Computational Biology, Centre for

DNA Fingerprinting & Diagnostics, Hyderabad, Andhra Pradesh 500 001, India.

The MICdb is a comprehensive relational database of perfect microsatellites

extracted from completely sequenced and annotated genomes of bacteria and

archaea. The current version MICdb3.0 is an updated and revised version of

MICdb2.0. As compared with the previous version MICdb2.0, the current release is

significantly improved in terms of much larger coverage of genomes, improved

presentation of queried results, user-friendly administration module to manage

Simple Sequence Repeat (SSR) data such as addition of new genomes, deletion of

obsolete data, etc., and also removal of certain features deemed to be redundant.

The new web-interface to the database called Microsatellite Analysis Server

(MICAS) version 3.0 has been improved by the addition of powerful high-quality

visualization tools to view the query results in the form of pie charts and bar

graphs. All the query results and graphs can be exported in different formats so

that the users can use them for further analysis. MICAS3.0 is also equipped with

a unique genome comparison module using which users can do pair-wise comparison

of genomes with regard to their microsatellite distribution. The advanced search

module can be used to filter the repeats based on certain criteria such as

filtering repeats of a particular motif/repeat size, extracting repeats of

coding/non-coding regions, sort repeats, etc. The MICdb database has, therefore,

been made portable to be administered by a person with the necessary

administrative privileges. The MICdb3.0 database and analysis server can be

accessed for free from www.cdfd.org.in/micas. Database URL:

http://www.cdfd.org.in/micas.

DOI: 10.1093/database/bau005

PMCID: PMC3926409

PMID: 24536078 [Indexed for MEDLINE]

896. JMIR Mhealth Uhealth. 2014 Feb 14;2(1):e5. doi: 10.2196/mhealth.3165.

Integrating mobile phones into medical abortion provision: intervention

development, use, and lessons learned from a randomized controlled trial.

de Tolly KM(1), Constant D.

Author information:

(1)Cell-Life, Cape Town, South Africa. kmdetolly@gmail.com.

BACKGROUND: Medical abortion is legal in South Africa but access and

acceptability are hampered by the current protocol requiring a follow-up visit to

assess abortion completion.

OBJECTIVE: To assess the feasibility and efficacy of information and follow-up

provided via mobile phone after medical abortion in a randomized controlled trial

(RCT).

METHODS: Mobile phones were used in three ways in the study: (1) coaching women

through medical abortion using short message service (SMS; text messages); (2) a

questionnaire to assess abortion completion via unstructured supplementary

service data (USSD, a protocol used by GSM mobile telephones that allows the user

to interact with a server via text-based menus) and the South African mobile

instant message and social networking application Mxit; and (3) family planning

information via SMS, mobisite and Mxit. A needs and context assessment was done

to learn about women's experiences undergoing medical abortion and their use of

mobile phones. After development, the mobile interventions were piloted.

Recruitment was done by field workers at the clinics. In the RCT, women were

interviewed at baseline and exit. Computer logs were also analyzed. All study

participants received standard of care at the clinics.

RESULTS: In the RCT, 234 women were randomized to the intervention group. Eight

did not receive the intervention due to invalid numbers, mis-registration, system

failure, or opt-out, leaving 226 participants receiving the full intervention. Of

the 226, 190 returned and were interviewed at their clinic follow-up visit. The

SMSs were highly acceptable, with 97.9% (186/190) saying that the SMSs helped

them through the medical abortion. In terms of mobile phone privacy, 86.3%

(202/234) said that it was not likely or possible that someone would see SMSs on

their phone, although at exit, 20% (38/190) indicated that they had worried about

phone privacy. Having been given training at baseline and subsequently asked via

SMS to complete the self-assessment questionnaire, 90.3% (204/226) attempted it,

and of those, 86.3% (176/204) reached an endpoint of the questionnaire. For the

family planning information, a preference for SMS was indicated by study clients,

although the publicly available Mxit/mobisite was heavily used (813,375 pages

were viewed) over the study duration.

CONCLUSIONS: SMS provided a good medium for timed, "push" information that guided

and supported women through medical abortion. Women were able to perform a

self-assessment questionnaire via mobile phones if provided training and prompted

by SMS. Phone privacy needs to be protected in similar settings. This study may

contribute to the successful expansion of medical abortion provision aided by

mobile phones.

TRIAL REGISTRATION: Pan African Clinical Trials Registry (PACTR):

PACTR201302000427144;

http://www.pactr.org/ATMWeb/appmanager/atm/atmregistry?dar=true&tNo=PACTR20130200

0427144 (Archived by WebCite at http://www.webcitation.org/6N0fnZfzm).

DOI: 10.2196/mhealth.3165

PMCID: PMC4114479

PMID: 25098569

897. PLoS One. 2014 Feb 5;9(2):e88222. doi: 10.1371/journal.pone.0088222. eCollection

2014.

Secondary structures of rRNAs from all three domains of life.

Petrov AS(1), Bernier CR(1), Gulen B(1), Waterbury CC(1), Hershkovits E(1), Hsiao

C(1), Harvey SC(1), Hud NV(1), Fox GE(2), Wartell RM(1), Williams LD(1).

Author information:

(1)Center for Ribosomal Origins and Evolution, Georgia Institute of Technology,

Atlanta, Georgia, United States of America ; School of Chemistry and Biochemistry

Georgia Institute of Technology, Atlanta, Georgia, United States of America.

(2)Center for Ribosomal Origins and Evolution, Georgia Institute of Technology,

Atlanta, Georgia, United States of America ; Department of Biology and

Biochemistry, University of Houston, Houston, Texas, United States of America.

Accurate secondary structures are important for understanding ribosomes, which

are extremely large and highly complex. Using 3D structures of ribosomes as

input, we have revised and corrected traditional secondary (2°) structures of

rRNAs. We identify helices by specific geometric and molecular interaction

criteria, not by co-variation. The structural approach allows us to incorporate

non-canonical base pairs on parity with Watson-Crick base pairs. The resulting

rRNA 2° structures are up-to-date and consistent with three-dimensional

structures, and are information-rich. These 2° structures are relatively simple

to understand and are amenable to reproduction and modification by end-users. The

2° structures made available here broadly sample the phylogenetic tree and are

mapped with a variety of data related to molecular interactions and geometry,

phylogeny and evolution. We have generated 2° structures for both large subunit

(LSU) 23S/28S and small subunit (SSU) 16S/18S rRNAs of Escherichia coli, Thermus

thermophilus, Haloarcula marismortui (LSU rRNA only), Saccharomyces cerevisiae,

Drosophila melanogaster, and Homo sapiens. We provide high-resolution editable

versions of the 2° structures in several file formats. For the SSU rRNA, the 2°

structures use an intuitive representation of the central pseudoknot where base

triples are presented as pairs of base pairs. Both LSU and SSU secondary maps are

available (http://apollo.chemistry.gatech.edu/RibosomeGallery). Mapping of data

onto 2° structures was performed on the RiboVision server

(http://apollo.chemistry.gatech.edu/RiboVision).

DOI: 10.1371/journal.pone.0088222

PMCID: PMC3914948

PMID: 24505437 [Indexed for MEDLINE]

898. Bioinformatics. 2014 Feb 1;30(3):335-42. doi: 10.1093/bioinformatics/btt691. Epub

2013 Nov 26.

mCSM: predicting the effects of mutations in proteins using graph-based

signatures.

Pires DE(1), Ascher DB, Blundell TL.

Author information:

(1)Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK and

ACRF Rational Drug Discovery Centre and Biota Structural Biology Laboratory, St

Vincents Institute of Medical Research, Fitzroy, VIC, 3065, Australia.

MOTIVATION: Mutations play fundamental roles in evolution by introducing

diversity into genomes. Missense mutations in structural genes may become either

selectively advantageous or disadvantageous to the organism by affecting protein

stability and/or interfering with interactions between partners. Thus, the

ability to predict the impact of mutations on protein stability and interactions

is of significant value, particularly in understanding the effects of Mendelian

and somatic mutations on the progression of disease. Here, we propose a novel

approach to the study of missense mutations, called mCSM, which relies on

graph-based signatures. These encode distance patterns between atoms and are used

to represent the protein residue environment and to train predictive models. To

understand the roles of mutations in disease, we have evaluated their impacts not

only on protein stability but also on protein-protein and protein-nucleic acid

interactions.

RESULTS: We show that mCSM performs as well as or better than other methods that

are used widely. The mCSM signatures were successfully used in different tasks

demonstrating that the impact of a mutation can be correlated with the

atomic-distance patterns surrounding an amino acid residue. We showed that mCSM

can predict stability changes of a wide range of mutations occurring in the

tumour suppressor protein p53, demonstrating the applicability of the proposed

method in a challenging disease scenario.

AVAILABILITY AND IMPLEMENTATION: A web server is available at

http://structure.bioc.cam.ac.uk/mcsm.

DOI: 10.1093/bioinformatics/btt691

PMCID: PMC3904523

PMID: 24281696 [Indexed for MEDLINE]

899. Bioinformatics. 2014 Feb 1;30(3):439-41. doi: 10.1093/bioinformatics/btt680. Epub

2013 Nov 22.

MemBuilder: a web-based graphical interface to build heterogeneously mixed

membrane bilayers for the GROMACS biomolecular simulation program.

Ghahremanpour MM(1), Arab SS, Aghazadeh SB, Zhang J, van der Spoel D.

Author information:

(1)Department of Bioinformatics, School of Computer Science, Institute for

Research in Fundamental Sciences (IPM), Tehran, Iran, Department of Biophysics,

School of Biological Sciences, Tarbiat Modares University, Tehran, Iran, School

of Mathematics, Institute for Research in Fundamental Sciences (IPM), Tehran,

Iran, Department of Computer Science, Faculty of Engineering, University of

Tehran, Tehran, Iran, Department of Cell and Molecular Biology, Uppsala

University, SE-75124 Uppsala, Sweden and Department of Chemistry, Zhejiang

University, Hangzhou 310027, China.

MOTIVATION: Molecular dynamics (MD) simulations have had a profound impact on

studies of membrane proteins during past two decades, but the accuracy of MD

simulations of membranes is limited by the quality of membrane models and the

applied force fields. Membrane models used in MD simulations mostly contain one

kind of lipid molecule. This is far from reality, for biological membranes always

contain more than one kind of lipid molecule. Moreover, the lipid composition and

their distribution are functionally important. As a result, there is a necessity

to prepare more realistic lipid membranes containing different types of lipids at

physiological concentrations.

RESULTS: To automate and simplify the building process of heterogeneous lipid

bilayers as well as providing molecular topologies for included lipids based on

both united and all-atom force fields, we provided MemBuilder as a web-based

graphical user interface.

AVAILABILITY AND IMPLEMENTATION: MemBuilder is a free web server available from

www.membuilder.org.

DOI: 10.1093/bioinformatics/btt680

PMID: 24273238 [Indexed for MEDLINE]

900. BMC Bioinformatics. 2014 Feb 1;15:36. doi: 10.1186/1471-2105-15-36.

uPEPperoni: an online tool for upstream open reading frame location and analysis

of transcript conservation.

Skarshewski A, Stanton-Cook M, Huber T, Al Mansoori S, Smith R, Beatson SA,

Rothnagel JA(1).

Author information:

(1)School of Chemistry and Molecular Biosciences, University of Queensland,

Brisbane, QLD 4072, Australia. j.rothnagel@uq.edu.au.

BACKGROUND: Several small open reading frames located within the 5' untranslated

regions of mRNAs have recently been shown to be translated. In humans, about 50%

of mRNAs contain at least one upstream open reading frame representing a large

resource of coding potential. We propose that some upstream open reading frames

encode peptides that are functional and contribute to proteome complexity in

humans and other organisms. We use the term uPEPs to describe peptides encoded by

upstream open reading frames.

RESULTS: We have developed an online tool, termed uPEPperoni, to facilitate the

identification of putative bioactive peptides. uPEPperoni detects conserved

upstream open reading frames in eukaryotic transcripts by comparing query

nucleotide sequences against mRNA sequences within the NCBI RefSeq database. The

algorithm first locates the main coding sequence and then searches for open

reading frames 5' to the main start codon which are subsequently analysed for

conservation. uPEPperoni also determines the substitution frequency for both the

upstream open reading frames and the main coding sequence. In addition, the

uPEPperoni tool produces sequence identity heatmaps which allow rapid visual

inspection of conserved regions in paired mRNAs.

CONCLUSIONS: uPEPperoni features user-nominated settings including, nucleotide

match/mismatch, gap penalties, Ka/Ks ratios and output mode. The heatmap output

shows levels of identity between any two sequences and provides easy recognition

of conserved regions. Furthermore, this web tool allows comparison of

evolutionary pressures acting on the upstream open reading frame against other

regions of the mRNA. Additionally, the heatmap web applet can also be used to

visualise the degree of conservation in any pair of sequences. uPEPperoni is

freely available on an interactive web server at

http://upep-scmb.biosci.uq.edu.au.

DOI: 10.1186/1471-2105-15-36

PMCID: PMC3914846

PMID: 24484385 [Indexed for MEDLINE]

901. Comput Biol Med. 2014 Feb;45:157-60. doi: 10.1016/j.compbiomed.2013.12.007. Epub

2013 Dec 21.

miRClassify: an advanced web server for miRNA family classification and

annotation.

Zou Q(1), Mao Y(2), Hu L(2), Wu Y(2), Ji Z(3).

Author information:

(1)School of Information Science and Technology, Xiamen University, Xiamen,

Fujian 361005, China. Electronic address: zouquan@xmu.edu.cn. (2)School of

Information Science and Technology, Xiamen University, Xiamen, Fujian 361005,

China. (3)State Key Laboratory of Stress Cell Biology, School of Life Sciences,

Xiamen University, Xiamen 361005, China; The Key Laboratory for Chemical Biology

of Fujian Province, Xiamen University, Xiamen 361005, China. Electronic address:

appo@xmu.edu.cn.

MicroRNA (miRNA) family is a group of miRNAs that derive from the common

ancestor. Normally, members from the same miRNA family have similar physiological

functions; however, they are not always conserved in primary sequence or

secondary structure. Proper family prediction from primary sequence will be

helpful for accurate identification and further functional annotation of novel

miRNA. Therefore, we introduced a novel machine learning-based web server, the

miRClassify, which can rapidly identify miRNA from the primary sequence and

classify it into a miRNA family regardless of similarity in sequence and

structure. Additionally, the medical implication of the miRNA family is also

provided when it is available in PubMed. The web server is accessible at the link

http://datamining.xmu.edu.cn/software/MIR/home.html.

Crown Copyright © 2013. Published by Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiomed.2013.12.007

PMID: 24480175 [Indexed for MEDLINE]

902. J Bioinform Comput Biol. 2014 Feb;12(1):1450005. doi: 10.1142/S021972001450005X.

Epub 2014 Jan 21.

PhD7Faster: predicting clones propagating faster from the Ph.D.-7 phage display

peptide library.

Ru B(1), 't Hoen PA, Nie F, Lin H, Guo FB, Huang J.

Author information:

(1)Center of Bioinformatics (COBI), Key Laboratory for NeuroInformation of

Ministry of Education, University of Electronic Science and Technology of China,

Chengdu 610054, P. R. China.

Phage display can rapidly discover peptides binding to any given target; thus, it

has been widely used in basic and applied research. Each round of panning

consists of two basic processes: Selection and amplification. However, recent

studies have showed that the amplification step would decrease the diversity of

phage display libraries due to different propagation capacity of phage clones.

This may induce phages with growth advantage rather than specific affinity to

appear in the final experimental results. The peptides displayed by such phages

are termed as propagation-related target-unrelated peptides (PrTUPs). They would

mislead further analysis and research if not removed. In this paper, we describe

PhD7Faster, an ensemble predictor based on support vector machine (SVM) for

predicting clones with growth advantage from the Ph.D.-7 phage display peptide

library. By using reduced dipeptide composition (ReDPC) as features, an accuracy

(Acc) of 79.67% and a Matthews correlation coefficient (MCC) of 0.595 were

achieved in 5-fold cross-validation. In addition, the SVM-based model was

demonstrated to perform better than several representative machine learning

algorithms. We anticipate that PhD7Faster can assist biologists to exclude

potential PrTUPs and accelerate the finding of specific binders from the popular

Ph.D.-7 library. The web server of PhD7Faster can be freely accessed at

http://immunet.cn/sarotup/cgi-bin/PhD7Faster.pl.

DOI: 10.1142/S021972001450005X

PMID: 24467763 [Indexed for MEDLINE]

903. J Bioinform Comput Biol. 2014 Feb;12(1):1450003. doi: 10.1142/S0219720014500036.

Epub 2014 Jan 7.

Accurate discrimination of outer membrane proteins using secondary structure

element alignment and support vector machine.

Ni Q(1), Zou L.

Author information:

(1)Department of Microbiology, College of Basic Medical Sciences, Third Military

Medical University, No. 30 Gaotanyan Road, Shapingba District, Chongqing 400038,

P. R. China.

Outer membrane proteins (OMPs) play critical roles in many cellular processes and

discriminating OMPs from other types of proteins is very important for OMPs

identification in bacterial genomic proteins. In this study, a method SSEA\_SVM is

developed using secondary structure element alignment and support vector machine.

Moreover, a novel kernel function is designed to utilize secondary structure

information in the support vector machine classifier. A benchmark dataset, which

consists of 208 OMPs, 673 globular proteins, and 206 α-helical membrane proteins,

is used to evaluate the performance of SSEA\_SVM. A high accuracy of 97.7% with

0.926 MCC is achieved while SSEA\_SVM is applied to discriminating OMPs and

non-OMPs. In comparison with existing methods in the literature, SSEA\_SVM is also

highly competitive. We suggest that SSEA\_SVM is a much more promising method to

identify OMPs in genomic proteins. A web server that implements SSEA\_SVM is

freely available at http://bioinfo.tmmu.edu.cn/SSEA\_SVM/.

DOI: 10.1142/S0219720014500036

PMID: 24467761 [Indexed for MEDLINE]

904. Proteins. 2014 Feb;82(2):250-67. doi: 10.1002/prot.24370. Epub 2013 Oct 17.

DockRank: ranking docked conformations using partner-specific sequence

homology-based protein interface prediction.

Xue LC(1), Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V.

Author information:

(1)Bioinformatics and Computational Biology program, Iowa State University, Ames,

Iowa.

Selecting near-native conformations from the immense number of conformations

generated by docking programs remains a major challenge in molecular docking. We

introduce DockRank, a novel approach to scoring docked conformations based on the

degree to which the interface residues of the docked conformation match a set of

predicted interface residues. DockRank uses interface residues predicted by

partner-specific sequence homology-based protein-protein interface predictor

(PS-HomPPI), which predicts the interface residues of a query protein with a

specific interaction partner. We compared the performance of DockRank with

several state-of-the-art docking scoring functions using Success Rate (the

percentage of cases that have at least one near-native conformation among the top

m conformations) and Hit Rate (the percentage of near-native conformations that

are included among the top m conformations). In cases where it is possible to

obtain partner-specific (PS) interface predictions from PS-HomPPI, DockRank

consistently outperforms both (i) ZRank and IRAD, two state-of-the-art

energy-based scoring functions (improving Success Rate by up to 4-fold); and (ii)

Variants of DockRank that use predicted interface residues obtained from several

protein interface predictors that do not take into account the binding partner in

making interface predictions (improving success rate by up to 39-fold). The

latter result underscores the importance of using partner-specific interface

residues in scoring docked conformations. We show that DockRank, when used to

re-rank the conformations returned by ClusPro, improves upon the original ClusPro

rankings in terms of both Success Rate and Hit Rate. DockRank is available as a

server at http://einstein.cs.iastate.edu/DockRank/.

Copyright © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24370

PMCID: PMC4417613

PMID: 23873600 [Indexed for MEDLINE]

905. Proteins. 2014 Feb;82(2):175-86. doi: 10.1002/prot.24299. Epub 2013 Oct 17.

Fragment-based modeling of membrane protein loops: successes, failures, and

prospects for the future.

Kelm S(1), Vangone A, Choi Y, Ebejer JP, Shi J, Deane CM.

Author information:

(1)Department of Statistics, University of Oxford, Oxford, United Kingdom.

Membrane proteins (MPs) have become a major focus in structure prediction, due to

their medical importance. There is, however, a lack of fast and reliable methods

that specialize in the modeling of MP loops. Often methods designed for soluble

proteins (SPs) are applied directly to MPs. In this article, we investigate the

validity of such an approach in the realm of fragment-based methods. We also

examined the differences in membrane and soluble protein loops that might affect

accuracy. We test our ability to predict soluble and MP loops with the previously

published method FREAD. We show that it is possible to predict accurately the

structure of MP loops using a database of MP fragments (0.5-1 Å median

root-mean-square deviation). The presence of homologous proteins in the database

helps prediction accuracy. However, even when homologues are removed better

results are still achieved using fragments of MPs (0.8-1.6 Å) rather than SPs

(1-4 Å) to model MP loops. We find that many fragments of SPs have shapes similar

to their MP counterparts but have very different sequences; however, they do not

appear to differ in their substitution patterns. Our findings may allow further

improvements to fragment-based loop modeling algorithms for MPs. The current

version of our proof-of-concept loop modeling protocol produces high-accuracy

loop models for MPs and is available as a web server at

http://medeller.info/fread.

Copyright © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24299

PMID: 23589399 [Indexed for MEDLINE]

906. Surg Innov. 2014 Feb;21(1):65-73. doi: 10.1177/1553350613484824. Epub 2013 Apr

16.

Magnetic compression in gastrointestinal and bilioenteric anastomosis: how much

force?

Lambe T(1), Ríordáin MG, Cahill RA, Cantillon-Murphy P.

Author information:

(1)1University College Cork, Cork, Ireland.

AIM: The concept of compression alimentary anastomosis is well established.

Recently, magnetic axial alignment pressures have been encompassed within such

device constructs. We quantify the magnetic compression force and pressure

required to successfully achieve gastrointestinal and bilioenteric anastomosis by

in-depth interrogation of the reported literature.

METHODS: Reports of successful deployment and proof of anastomotic patency on

survival were scrutinized to quantify the necessary dimensions and strengths of

magnetic devices in (a) gastroenteral anastomosis in live porcine models and (b)

bilioenteric anastomosis in the clinical setting. Using a calculatory tool

developed for this work (magnetic force determination algorithm, MAGDA), ideal

magnetic force and compression pressure were quantified from successful reports

with regard to their variance by intermagnet separation.

RESULTS: Optimized ranges for both compression force and pressure were determined

for successful porcine gastroenteral and clinical bilioenteric anastomoses. For

gastroenteral anastomoses (porcine investigations), an optimized compression

force between 2.55 and 3.57 kg at 2-mm intermagnet separation is recommended. The

associated compression pressure should not exceed 60 N/cm(2). Successful

bilioenteric anastomoses is best clinically achieved with intermagnet compression

of 18 to 31 g and associated pressures between 1 and 3.5 N/mm(2) (at 2-mm

intermagnet separation).

CONCLUSION: The creation of magnetic compression anastomoses using permanent

magnets demonstrates a remarkable resilience to variations in magnetic force and

pressure exertion. However, inappropriate selection of compression

characteristics and magnet dimensions may incur difficulties. Recommendations of

this work and the availability of the free online tool (http://magda.ucc.ie/) may

facilitate a factor of robustness in the design and refinement of future devices.

DOI: 10.1177/1553350613484824

PMID: 23592733 [Indexed for MEDLINE]

907. Bioinformation. 2014 Jan 29;10(1):48-51. doi: 10.6026/97320630010048. eCollection

2014.

PPS: A computing engine to find Palindromes in all Protein sequences.

Ahmed Z(1), Gurusaran M(1), Narayana P(1), Kumar KS(1), Mohanapriya J(1),

Vaishnavi MK(1), Sekar K(1).

Author information:

(1)Supercomputer Education and Research Centre, Indian Institute of Science,

Bangalore 560 012, India.

The primary structure of a protein molecule comprises a linear chain of amino

acid residues. Certain parts of this linear chain are unique in nature and

function. They can be classified under different categories and their roles

studied in detail. Two such unique categories are the palindromic sequences and

the Single Amino Acid Repeats (SAARs), which plays a major role in the structure,

function and evolution of the protein molecule. In spite of their presence in

various protein sequences, palindromes have not yet been investigated in detail.

Thus, to enable a comprehensive understanding of these sequences, a computing

engine, PPS, has been developed. The users can search the occurrences of

palindromes and SAARs in all the protein sequences available in various databases

and can view the three-dimensional structures (in case it is available in the

known three-dimensional protein structures deposited to the Protein Data Bank)

using the graphics plug-in Jmol. The proposed server is the first of its kind and

can be freely accessed through the World Wide Web.AVAILABILITY: URL

http://pranag.physics.iisc.ernet.in/pps/

DOI: 10.6026/97320630010048

PMCID: PMC3916820

PMID: 24516327

908. BMC Res Notes. 2014 Jan 27;7:63. doi: 10.1186/1756-0500-7-63.

Support vector machine (SVM) based multiclass prediction with basic statistical

analysis of plasminogen activators.

Muthukrishnan S, Puri M(1), Lefevre C.

Author information:

(1)Fermentation and Protein Biotechnology Laboratory, Department of

Biotechnology, Punjabi University, Patiala, India, 2CSIR-IMTECH, Chandigarh,

India. munish.puri@deakin.edu.au.

BACKGROUND: Plasminogen (Pg), the precursor of the proteolytic and fibrinolytic

enzyme of blood, is converted to the active enzyme plasmin (Pm) by different

plasminogen activators (tissue plasminogen activators and urokinase), including

the bacterial activators streptokinase and staphylokinase, which activate Pg to

Pm and thus are used clinically for thrombolysis. The identification of

Pg-activators is therefore an important step in understanding their functional

mechanism and derives new therapies.

METHODS: In this study, different computational methods for predicting

plasminogen activator peptide sequences with high accuracy were investigated,

including support vector machines (SVM) based on amino acid (AC), dipeptide

composition (DC), PSSM profile and Hybrid methods used to predict different

Pg-activators from both prokaryotic and eukaryotic origins.

RESULTS: Overall maximum accuracy, evaluated using the five-fold cross validation

technique, was 88.37%, 84.32%, 87.61%, 85.63% in 0.87, 0.83,0.86 and 0.85 MCC

with amino (AC) or dipeptide composition (DC), PSSM profile and Hybrid methods

respectively. Through this study, we have found that the different subfamilies of

Pg-activators are quite closely correlated in terms of amino, dipeptide, PSSM and

Hybrid compositions. Therefore, our prediction results show that plasminogen

activators are predictable with a high accuracy from their primary sequence.

Prediction performance was also cross-checked by confusion matrix and ROC

(Receiver operating characteristics) analysis. A web server to facilitate the

prediction of Pg-activators from primary sequence data was implemented.

CONCLUSION: The results show that dipeptide, PSSM profile, and Hybrid based

methods perform better than single amino acid composition (AC). Furthermore, we

also have developed a web server, which predicts the Pg-activators and their

classification (available online at

http://mamsap.it.deakin.edu.au/plas\_pred/home.html). Our experimental results

show that our approaches are faster and achieve generally a good prediction

performance.

DOI: 10.1186/1756-0500-7-63

PMCID: PMC3924408

PMID: 24468032 [Indexed for MEDLINE]

909. Int J Mol Sci. 2014 Jan 24;15(2):1746-66. doi: 10.3390/ijms15021746.

iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and

pseudo amino acid components.

Qiu WR(1), Xiao X(2), Chou KC(3).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen 333046, China.

qiuone@163.com. (2)Computer Department, Jing-De-Zhen Ceramic Institute,

Jingdezhen 333046, China. xxiao@gordonlifescience.org. (3)Center of Excellence in

Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi

Arabia. kcchou@gordonlifescience.org.

Meiosis and recombination are the two opposite aspects that coexist in a DNA

system. As a driving force for evolution by generating natural genetic

variations, meiotic recombination plays a very important role in the formation of

eggs and sperm. Interestingly, the recombination does not occur randomly across a

genome, but with higher probability in some genomic regions called "hotspots",

while with lower probability in so-called "coldspots". With the ever-increasing

amount of genome sequence data in the postgenomic era, computational methods for

effectively identifying the hotspots and coldspots have become urgent as they can

timely provide us with useful insights into the mechanism of meiotic

recombination and the process of genome evolution as well. To meet the need, we

developed a new predictor called "iRSpot-TNCPseAAC", in which a DNA sample was

formulated by combining its trinucleotide composition (TNC) and the pseudo amino

acid components (PseAAC) of the protein translated from the DNA sample according

to its genetic codes. The former was used to incorporate its local or short-rage

sequence order information; while the latter, its global and long-range one.

Compared with the best existing predictor in this area, iRSpot-TNCPseAAC achieved

higher rates in accuracy, Mathew's correlation coefficient, and sensitivity,

indicating that the new predictor may become a useful tool for identifying the

recombination hotspots and coldspots, or, at least, become a complementary tool

to the existing methods. It has not escaped our notice that the aforementioned

novel approach to incorporate the DNA sequence order information into a discrete

model may also be used for many other genome analysis problems. The web-server

for iRSpot-TNCPseAAC is available at http://www.jci-bioinfo.cn/iRSpot-TNCPseAAC.

Furthermore, for the convenience of the vast majority of experimental scientists,

a step-by-step guide is provided on how to use the current web server to obtain

their desired result without the need to follow the complicated mathematical

equations.

DOI: 10.3390/ijms15021746

PMCID: PMC3958819

PMID: 24469313 [Indexed for MEDLINE]

910. Genome Biol. 2014 Jan 22;15(1):R18. doi: 10.1186/gb-2014-15-1-r18.

PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis.

Chen B, Yun J, Kim MS, Mendell JT, Xie Y.

CLIP-seq is widely used to study genome-wide interactions between RNA-binding

proteins and RNAs. However, there are few tools available to analyze CLIP-seq

data, thus creating a bottleneck to the implementation of this methodology. Here,

we present PIPE-CLIP, a Galaxy framework-based comprehensive online pipeline for

reliable analysis of data generated by three types of CLIP-seq protocol:

HITS-CLIP, PAR-CLIP and iCLIP. PIPE-CLIP provides both data processing and

statistical analysis to determine candidate cross-linking regions, which are

comparable to those regions identified from the original studies or using

existing computational tools. PIPE-CLIP is available at

http://pipeclip.qbrc.org/.

DOI: 10.1186/gb-2014-15-1-r18

PMCID: PMC4054095

PMID: 24451213 [Indexed for MEDLINE]

911. F1000Res. 2014 Jan 21;3:17. doi: 10.12688/f1000research.3-17.v1. eCollection

2014.

Ranking the quality of protein structure models using sidechain based network

properties.

Ghosh S(1), Vishveshwara S(2).

Author information:

(1)Molecular Biophysics Unit, Indian Institute of Science, Bangalore, 560012,

India ; I.I.Sc. Mathematics Initiative, Indian Institute of Science, Bangalore,

560012, India. (2)I.I.Sc. Mathematics Initiative, Indian Institute of Science,

Bangalore, 560012, India.

Determining the correct structure of a protein given its sequence still remains

an arduous task with many researchers working towards this goal. Most structure

prediction methodologies result in the generation of a large number of probable

candidates with the final challenge being to select the best amongst these. In

this work, we have used Protein Structure Networks of native and modeled proteins

in combination with Support Vector Machines to estimate the quality of a protein

structure model and finally to provide ranks for these models. Model ranking is

performed using regression analysis and helps in model selection from a group of

many similar and good quality structures. Our results show that structures with a

rank greater than 16 exhibit native protein-like properties while those below 10

are non-native like. The tool is also made available as a web-server (

http://vishgraph.mbu.iisc.ernet.in/GraProStr/native\_non\_native\_ranking.html),

where, 5 modelled structures can be evaluated at a given time.

DOI: 10.12688/f1000research.3-17.v1

PMCID: PMC4038323

PMID: 25580218

912. Bioinformatics. 2014 Jan 15;30(2):234-41. doi: 10.1093/bioinformatics/btt642.

Epub 2013 Nov 8.

PTMTreeSearch: a novel two-stage tree-search algorithm with pruning rules for the

identification of post-translational modification of proteins in MS/MS spectra.

Kertész-Farkas A(1), Reiz B, Vera R, Myers MP, Pongor S.

Author information:

(1)Protein Structure and Bioinformatics Group, International Centre for Genetic

Engineering and Biotechnology, AREA Research Park, 99 Padriciano, Trieste, Italy,

34149, Institute of Biophysics, Biological Research Centre, Temesvari krt. 62,

H-6727 Szeged, Hungary, Protein Networks Group, International Centre for Genetic

Engineering and Biotechnology, AREA Research Park, Padriciano 99, 34149 Trieste,

Italy and Faculty of Information Technology, Pázmány Péter Catholic University,

Práter u. 50/a, H-1083 Budapest, Hungary.

MOTIVATION: Tandem mass spectrometry has become a standard tool for identifying

post-translational modifications (PTMs) of proteins. Algorithmic searches for

PTMs from tandem mass spectrum data (MS/MS) tend to be hampered by noisy data as

well as by a combinatorial explosion of search space. This leads to high

uncertainty and long search-execution times.

RESULTS: To address this issue, we present PTMTreeSearch, a new algorithm that

uses a large database of known PTMs to identify PTMs from MS/MS data. For a given

peptide sequence, PTMTreeSearch builds a computational tree wherein each path

from the root to the leaves is labeled with the amino acids of a peptide

sequence. Branches then represent PTMs. Various empirical tree pruning rules have

been designed to decrease the search-execution time by eliminating biologically

unlikely solutions. PTMTreeSearch first identifies a relatively small set of high

confidence PTM types, and in a second stage, performs a more exhaustive search on

this restricted set using relaxed search parameter settings. An analysis of

experimental data shows that using the same criteria for false discovery,

PTMTreeSearch annotates more peptides than the current state-of-the-art methods

and PTM identification algorithms, and achieves this at roughly the same

execution time. PTMTreeSearch is implemented as a plugable scoring function in

the X!Tandem search engine.

AVAILABILITY: The source code of PTMTreeSearch and a demo server application can

be found at http://net.icgeb.org/ptmtreesearch

DOI: 10.1093/bioinformatics/btt642

PMID: 24215026 [Indexed for MEDLINE]

913. J Neurol Sci. 2014 Jan 15;336(1-2):113-5. doi: 10.1016/j.jns.2013.10.019. Epub

2013 Oct 17.

Validation of a seizure-related injury model.

Tiamkao S(1), Kaewkiow N(2), Pranbul S(3), Sawanyawisuth K(4); Integrated

Epilepsy Research group.

Author information:

(1)Department of Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen

40002, Thailand; Integrated Epilepsy Research Group, Khon Kaen University,

Thailand. Electronic address: somtia@kku.ac.th. (2)Department of Medicine,

Faculty of Medicine, Khon Kaen University, Khon Kaen 40002, Thailand.

(3)Integrated Epilepsy Research Group, Khon Kaen University, Thailand; Nursing

Division, Srinagarind Hospital, Faculty of Medicine, Khon Kaen University, Khon

Kaen 40002, Thailand. (4)Department of Medicine, Faculty of Medicine, Khon Kaen

University, Khon Kaen 40002, Thailand; Researches and Diagnostic Center for

Emerging Infectious Diseases, Faculty of Medicine, Khon Kaen University, Khon

Kaen, Thailand.

OBJECTIVE: Persons with epilepsy (PWEs) are more prone to accidents than healthy

people. A previous study provided an online tool to predict the risk of

seizure-related injury (SRI) in individual PWEs. There is, however, no validation

of the formula.

METHODS: This is a cross-sectional study conducted in 10 community hospitals in

Thailand. PWEs with an age of over 18 years were enrolled and defined by having

had a seizure related injury (SRI). The probability of individual PWEs having a

SRI was calculated by the online tool (http://sribykku.webs.com). The probability

of this happening in all patients was calculated for sensitivity and specificity

when compared with real data.

RESULTS: There were 316 patients enrolled in the study. Of those, 122 patients

(38.6%) had a SRI. The sensitivity and specificity of having a SRI by the online

formula were 93.44% and 43.30%, respectively.

CONCLUSION: The online formula to predict SRI in PWEs is valid and provided

comparable sensitivity and specificity with a previous study that was conducted

in the tertiary care hospital.

© 2013.

DOI: 10.1016/j.jns.2013.10.019

PMID: 24209899 [Indexed for MEDLINE]

914. PLoS One. 2014 Jan 15;9(1):e84598. doi: 10.1371/journal.pone.0084598. eCollection

2014.

Deriving a mutation index of carcinogenicity using protein structure and protein

interfaces.

Espinosa O(1), Mitsopoulos K(1), Hakas J(1), Pearl F(2), Zvelebil M(1).

Author information:

(1)Breakthrough Breast Cancer Research Centre, Institute of Cancer Research,

London, United Kingdom. (2)UK Cancer Therapeutics Unit, The Institute of Cancer

Research, London, United Kingdom ; Translational Drug Discovery Group, School of

Life Sciences, University of Sussex, Brighton, United Kingdom.

With the advent of Next Generation Sequencing the identification of mutations in

the genomes of healthy and diseased tissues has become commonplace. While much

progress has been made to elucidate the aetiology of disease processes in cancer,

the contributions to disease that many individual mutations make remain to be

characterised and their downstream consequences on cancer phenotypes remain to be

understood. Missense mutations commonly occur in cancers and their consequences

remain challenging to predict. However, this knowledge is becoming more vital,

for both assessing disease progression and for stratifying drug treatment

regimes. Coupled with structural data, comprehensive genomic databases of

mutations such as the 1000 Genomes project and COSMIC give an opportunity to

investigate general principles of how cancer mutations disrupt proteins and their

interactions at the molecular and network level. We describe a comprehensive

comparison of cancer and neutral missense mutations; by combining features

derived from structural and interface properties we have developed a

carcinogenicity predictor, InCa (Index of Carcinogenicity). Upon comparison with

other methods, we observe that InCa can predict mutations that might not be

detected by other methods. We also discuss general limitations shared by all

predictors that attempt to predict driver mutations and discuss how this could

impact high-throughput predictions. A web interface to a server implementation is

publicly available at http://inca.icr.ac.uk/.

DOI: 10.1371/journal.pone.0084598

PMCID: PMC3893166

PMID: 24454733 [Indexed for MEDLINE]

915. BMC Bioinformatics. 2014 Jan 14;15:12. doi: 10.1186/1471-2105-15-12.

GIANT: pattern analysis of molecular interactions in 3D structures of

protein-small ligand complexes.

Kasahara K, Kinoshita K(1).

Author information:

(1)Graduate School of Information Sciences, Tohoku University, 6-3-09 Aoba,

Aramaki, Aoba-ku, Sendai, Miyagi 980-8597, Japan. kengo@ecei.tohoku.ac.jp.

BACKGROUND: Interpretation of binding modes of protein-small ligand complexes

from 3D structure data is essential for understanding selective ligand

recognition by proteins. It is often performed by visual inspection and sometimes

largely depends on a priori knowledge about typical interactions such as hydrogen

bonds and π-π stacking. Because it can introduce some biases due to scientists'

subjective perspectives, more objective viewpoints considering a wide range of

interactions are required.

DESCRIPTION: In this paper, we present a web server for analyzing protein-small

ligand interactions on the basis of patterns of atomic contacts, or "interaction

patterns" obtained from the statistical analyses of 3D structures of

protein-ligand complexes in our previous study. This server can guide visual

inspection by providing information about interaction patterns for each atomic

contact in 3D structures. Users can visually investigate what atomic contacts in

user-specified 3D structures of protein-small ligand complexes are statistically

overrepresented. This server consists of two main components: "Complex Analyzer",

and "Pattern Viewer". The former provides a 3D structure viewer with annotations

of interacting amino acid residues, ligand atoms, and interacting pairs of these.

In the annotations of interacting pairs, assignment to an interaction pattern of

each contact and statistical preferences of the patterns are presented. The

"Pattern Viewer" provides details of each interaction pattern. Users can see

visual representations of probability density functions of interactions, and a

list of protein-ligand complexes showing similar interactions.

CONCLUSIONS: Users can interactively analyze protein-small ligand binding modes

with statistically determined interaction patterns rather than relying on a

priori knowledge of the users, by using our new web server named GIANT that is

freely available at http://giant.hgc.jp/.

DOI: 10.1186/1471-2105-15-12

PMCID: PMC3897944

PMID: 24423161 [Indexed for MEDLINE]

916. BMC Bioinformatics. 2014 Jan 13;15:7. doi: 10.1186/1471-2105-15-7.

Skylign: a tool for creating informative, interactive logos representing sequence

alignments and profile hidden Markov models.

Wheeler TJ(1), Clements J, Finn RD.

Author information:

(1)HHMI Janelia Farm Research Campus, Ashburn, VA 20147, USA.

wheelert@janelia.hhmi.org.

BACKGROUND: Logos are commonly used in molecular biology to provide a compact

graphical representation of the conservation pattern of a set of sequences. They

render the information contained in sequence alignments or profile hidden Markov

models by drawing a stack of letters for each position, where the height of the

stack corresponds to the conservation at that position, and the height of each

letter within a stack depends on the frequency of that letter at that position.

RESULTS: We present a new tool and web server, called Skylign, which provides a

unified framework for creating logos for both sequence alignments and profile

hidden Markov models. In addition to static image files, Skylign creates a novel

interactive logo plot for inclusion in web pages. These interactive logos enable

scrolling, zooming, and inspection of underlying values. Skylign can avoid

sampling bias in sequence alignments by down-weighting redundant sequences and by

combining observed counts with informed priors. It also simplifies the

representation of gap parameters, and can optionally scale letter heights based

on alternate calculations of the conservation of a position.

CONCLUSION: Skylign is available as a website, a scriptable web service with a

RESTful interface, and as a software package for download. Skylign's interactive

logos are easily incorporated into a web page with just a few lines of HTML

markup. Skylign may be found at http://skylign.org.

DOI: 10.1186/1471-2105-15-7

PMCID: PMC3893531

PMID: 24410852 [Indexed for MEDLINE]

917. J Proteome Res. 2014 Jan 3;13(1):76-83. doi: 10.1021/pr400794x. Epub 2013 Dec 13.

Protannotator: a semiautomated pipeline for chromosome-wise functional annotation

of the "missing" human proteome.

Islam MT(1), Garg G, Hancock WS, Risk BA, Baker MS, Ranganathan S.

Author information:

(1)Department of Chemistry and Biomolecular Sciences and ‡ARC Centre of

Excellence in Bioinformatics, Macquarie University , Sydney, NSW 2109, Australia.

The chromosome-centric human proteome project (C-HPP) aims to define the complete

set of proteins encoded in each human chromosome. The neXtProt database

(September 2013) lists 20,128 proteins for the human proteome, of which 3831

human proteins (∼19%) are considered "missing" according to the standard metrics

table (released September 27, 2013). In support of the C-HPP initiative, we have

extended the annotation strategy developed for human chromosome 7 "missing"

proteins into a semiautomated pipeline to functionally annotate the "missing"

human proteome. This pipeline integrates a suite of bioinformatics analysis and

annotation software tools to identify homologues and map putative functional

signatures, gene ontology, and biochemical pathways. From sequential BLAST

searches, we have primarily identified homologues from reviewed nonhuman

mammalian proteins with protein evidence for 1271 (33.2%) "missing" proteins,

followed by 703 (18.4%) homologues from reviewed nonhuman mammalian proteins and

subsequently 564 (14.7%) homologues from reviewed human proteins. Functional

annotations for 1945 (50.8%) "missing" proteins were also determined. To

accelerate the identification of "missing" proteins from proteomics studies, we

generated proteotypic peptides in silico. Matching these proteotypic peptides to

ENCODE proteogenomic data resulted in proteomic evidence for 107 (2.8%) of the

3831 "missing proteins, while evidence from a recent membrane proteomic study

supported the existence for another 15 "missing" proteins. The chromosome-wise

functional annotation of all "missing" proteins is freely available to the

scientific community through our web server (http://biolinfo.org/protannotator).

DOI: 10.1021/pr400794x

PMID: 24313344 [Indexed for MEDLINE]

918. Adv Protein Chem Struct Biol. 2014;96:267-84. doi: 10.1016/bs.apcsb.2014.06.004.

Epub 2014 Aug 24.

High-resolution modeling of protein structures based on flexible fitting of

low-resolution structural data.

Zheng W(1), Tekpinar M(2).

Author information:

(1)Department of Physics, University at Buffalo, Buffalo, New York, USA.

Electronic address: wjzheng@buffalo.edu. (2)Department of Physics, Yuzuncu Yil

University, Kampus, Turkey.

To circumvent the difficulty of directly solving high-resolution biomolecular

structures, low-resolution structural data from Cryo-electron microscopy (EM) and

small angle solution X-ray scattering (SAXS) are increasingly used to explore

multiple conformational states of biomolecular assemblies. One promising venue to

obtain high-resolution structural models from low-resolution data is via

data-constrained flexible fitting. To this end, we have developed a new method

based on a coarse-grained Cα-only protein representation, and a modified form of

the elastic network model (ENM) that allows large-scale conformational changes

while maintaining the integrity of local structures including pseudo-bonds and

secondary structures. Our method minimizes a pseudo-energy which linearly

combines various terms of the modified ENM energy with an EM/SAXS-fitting score

and a collision energy that penalizes steric collisions. Unlike some previous

flexible fitting efforts using the lowest few normal modes, our method

effectively utilizes all normal modes so that both global and local structural

changes can be fully modeled with accuracy. This method is also highly efficient

in computing time. We have demonstrated our method using adenylate kinase as a

test case which undergoes a large open-to-close conformational change. The

EM-fitting method is available at a web server (http://enm.lobos.nih.gov), and

the SAXS-fitting method is available as a pre-compiled executable upon request.

© 2014 Elsevier Inc. All rights reserved.

DOI: 10.1016/bs.apcsb.2014.06.004

PMID: 25443961 [Indexed for MEDLINE]

919. Bioinformatics. 2014 Jan 1;30(1):121-2. doi: 10.1093/bioinformatics/btt614. Epub

2013 Nov 21.

Sequence alignment visualization in HTML5 without Java.

Gille C(1), Birgit W, Gille A.

Author information:

(1)Universitätsmedizin Berlin, Charité, Institut für Biochemie, 10117 Berlin,

Germany, Department of Plastic, Hannover Medical School, Hand and Reconstructive

Surgery, 30625 Hannover, Germany and Universität Heidelberg, Institut für

Experimentelle und Klinische Pharmakologie und Toxikologie, 68169 Mannheim,

Germany.

MOTIVATION: Java has been extensively used for the visualization of biological

data in the web. However, the Java runtime environment is an additional layer of

software with an own set of technical problems and security risks. HTML in its

new version 5 provides features that for some tasks may render Java unnecessary.

RESULTS: Alignment-To-HTML is the first HTML-based interactive visualization for

annotated multiple sequence alignments. The server side script interpreter can

perform all tasks like (i) sequence retrieval, (ii) alignment computation, (iii)

rendering, (iv) identification of a homologous structural models and (v)

communication with BioDAS-servers. The rendered alignment can be included in web

pages and is displayed in all browsers on all platforms including touch screen

tablets. The functionality of the user interface is similar to legacy Java

applets and includes color schemes, highlighting of conserved and variable

alignment positions, row reordering by drag and drop, interlinked 3D

visualization and sequence groups. Novel features are (i) support for multiple

overlapping residue annotations, such as chemical modifications, single

nucleotide polymorphisms and mutations, (ii) mechanisms to quickly hide residue

annotations, (iii) export to MS-Word and (iv) sequence icons.

CONCLUSION: Alignment-To-HTML, the first interactive alignment visualization that

runs in web browsers without additional software, confirms that to some extend

HTML5 is already sufficient to display complex biological data. The low speed at

which programs are executed in browsers is still the main obstacle. Nevertheless,

we envision an increased use of HTML and JavaScript for interactive biological

software.

AVAILABILITY AND IMPLEMENTATION: Under GPL at:

http://www.bioinformatics.org/strap/toHTML/.

DOI: 10.1093/bioinformatics/btt614

PMID: 24273246 [Indexed for MEDLINE]

920. BMC Bioinformatics. 2014;15 Suppl 8:S3. doi: 10.1186/1471-2105-15-S8-S3. Epub

2014 Jul 14.

Template-based C8-SCORPION: a protein 8-state secondary structure prediction

method using structural information and context-based features.

Yaseen A, Li Y.

BACKGROUND: Secondary structures prediction of proteins is important to many

protein structure modeling applications. Correct prediction of secondary

structures can significantly reduce the degrees of freedom in protein tertiary

structure modeling and therefore reduces the difficulty of obtaining high

resolution 3D models.

METHODS: In this work, we investigate a template-based approach to enhance

8-state secondary structure prediction accuracy. We construct structural

templates from known protein structures with certain sequence similarity. The

structural templates are then incorporated as features with sequence and

evolutionary information to train two-stage neural networks. In case of

structural templates absence, heuristic structural information is incorporated

instead.

RESULTS: After applying the template-based 8-state secondary structure prediction

method, the 7-fold cross-validated Q8 accuracy is 78.85%. Even templates from

structures with only 20%~30% sequence similarity can help improve the 8-state

prediction accuracy. More importantly, when good templates are available, the

prediction accuracy of less frequent secondary structures, such as 3-10 helices,

turns, and bends, are highly improved, which are useful for practical

applications.

CONCLUSIONS: Our computational results show that the templates containing

structural information are effective features to enhance 8-state secondary

structure predictions. Our prediction algorithm is implemented on a web server

named "C8-SCORPION" available at: http://hpcr.cs.odu.edu/c8scorpion.

DOI: 10.1186/1471-2105-15-S8-S3

PMCID: PMC4120151

PMID: 25080939 [Indexed for MEDLINE]

921. BMC Bioinformatics. 2014;15 Suppl 1:S8. doi: 10.1186/1471-2105-15-S1-S8. Epub

2014 Jan 10.

AnnotateGenomicRegions: a web application.

Zammataro L, DeMolfetta R, Bucci G, Ceol A, Muller H.

BACKGROUND: Modern genomic technologies produce large amounts of data that can be

mapped to specific regions in the genome. Among the first steps in interpreting

the results is annotation of genomic regions with known features such as genes,

promoters, CpG islands etc. Several tools have been published to perform this

task. However, using these tools often requires a significant amount of

bioinformatics skills and/or downloading and installing dedicated software.

RESULTS: Here we present AnnotateGenomicRegions, a web application that accepts

genomic regions as input and outputs a selection of overlapping and/or

neighboring genome annotations. Supported organisms include human (hg18, hg19),

mouse (mm8, mm9, mm10), zebrafish (danRer7), and Saccharomyces cerevisiae

(sacCer2, sacCer3). AnnotateGenomicRegions is accessible online on a public

server or can be installed locally. Some frequently used annotations and genomes

are embedded in the application while custom annotations may be added by the

user.

CONCLUSIONS: The increasing spread of genomic technologies generates the need for

a simple-to-use annotation tool for genomic regions that can be used by

biologists and bioinformaticians alike. AnnotateGenomicRegions meets this demand.

AnnotateGenomicRegions is an open-source web application that can be installed on

any personal computer or institute server. AnnotateGenomicRegions is available

at: http://cru.genomics.iit.it/AnnotateGenomicRegions.

DOI: 10.1186/1471-2105-15-S1-S8

PMCID: PMC4015944

PMID: 24564446 [Indexed for MEDLINE]

922. BMC Genomics. 2014;15 Suppl 4:S7. doi: 10.1186/1471-2164-15-S4-S7. Epub 2014 May

20.

NeEMO: a method using residue interaction networks to improve prediction of

protein stability upon mutation.

Giollo M, Martin AJ, Walsh I, Ferrari C, Tosatto SC.

BACKGROUND: The rapid growth of un-annotated missense variants poses challenges

requiring novel strategies for their interpretation. From the thermodynamic point

of view, amino acid changes can lead to a change in the internal energy of a

protein and induce structural rearrangements. This is of great relevance for the

study of diseases and protein design, justifying the development of prediction

methods for variant-induced stability changes.

RESULTS: Here we propose NeEMO, a tool for the evaluation of stability changes

using an effective representation of proteins based on residue interaction

networks (RINs). RINs are used to extract useful features describing interactions

of the mutant amino acid with its structural environment. Benchmarking shows

NeEMO to be very effective, allowing reliable predictions in different parts of

the protein such as β-strands and buried residues. Validation on a previously

published independent dataset shows that NeEMO has a Pearson correlation

coefficient of 0.77 and a standard error of 1 Kcal/mol, outperforming nine recent

methods. The NeEMO web server can be freely accessed from URL:

http://protein.bio.unipd.it/neemo/.

CONCLUSIONS: NeEMO offers an innovative and reliable tool for the annotation of

amino acid changes. A key contribution are RINs, which can be used for modeling

proteins and their interactions effectively. Interestingly, the approach is very

general, and can motivate the development of a new family of RIN-based protein

structure analyzers. NeEMO may suggest innovative strategies for bioinformatics

tools beyond protein stability prediction.

DOI: 10.1186/1471-2164-15-S4-S7

PMCID: PMC4083412

PMID: 25057121 [Indexed for MEDLINE]

923. Curr Top Med Chem. 2014;14(11):1399-415.

Tuning HERG out: antitarget QSAR models for drug development.

Braga RC, Alves VM, Silva MF, Muratov E, Fourches D, Tropsha A, Andrade CH(1).

Author information:

(1)LabMol, Faculdade de Farmacia, Universidade Federal de Goias, Rua 240, Qd. 87,

Setor Leste Universitario, Goiania, Goias 74605-170, Brazil. carolina@ufg.br.

Several non-cardiovascular drugs have been withdrawn from the market due to their

inhibition of hERG K+ channels that can potentially lead to severe heart

arrhythmia and death. As hERG safety testing is a mandatory FDArequired

procedure, there is a considerable interest for developing predictive

computational tools to identify and filter out potential hERG blockers early in

the drug discovery process. In this study, we aimed to generate predictive and

well-characterized quantitative structure-activity relationship (QSAR) models for

hERG blockage using the largest publicly available dataset of 11,958 compounds

from the ChEMBL database. The models have been developed and validated according

to OECD guidelines using four types of descriptors and four different

machine-learning techniques. The classification accuracies discriminating

blockers from non-blockers were as high as 0.83-0.93 on external set. Model

interpretation revealed several SAR rules, which can guide structural

optimization of some hERG blockers into non-blockers. We have also applied the

generated models for screening the World Drug Index (WDI) database and identify

putative hERG blockers and non-blockers among currently marketed drugs. The

developed models can reliably identify blockers and non-blockers, which could be

useful for the scientific community. A freely accessible web server has been

developed allowing users to identify putative hERG blockers and non-blockers in

chemical libraries of their interest (http://labmol.farmacia.ufg.br/predherg).

PMCID: PMC4593700

PMID: 24805060 [Indexed for MEDLINE]

924. Genome Biol Evol. 2014 Jan;6(1):12-33. doi: 10.1093/gbe/evt200.

Shared subgenome dominance following polyploidization explains grass genome

evolutionary plasticity from a seven protochromosome ancestor with 16K

protogenes.

Murat F(1), Zhang R, Guizard S, Flores R, Armero A, Pont C, Steinbach D,

Quesneville H, Cooke R, Salse J.

Author information:

(1)INRA/UBP UMR 1095 GDEC (Génétique, Diversité et Ecophysiologie des Céréales),

Clermont Ferrand, France.

Modern plant genomes are diploidized paleopolyploids. We revisited grass genome

paleohistory in response to the diploidization process through a detailed

investigation of the evolutionary fate of duplicated blocks. Ancestrally

duplicated genes can be conserved, deleted, and shuffled, defining dominant (bias

toward duplicate retention) and sensitive (bias toward duplicate erosion)

chromosomal fragments. We propose a new grass genome paleohistory deriving from

an ancestral karyotype structured in seven protochromosomes containing 16,464

protogenes and following evolutionary rules where 1) ancestral shared

polyploidizations shaped conserved dominant (D) and sensitive (S) subgenomes, 2)

subgenome dominance is revealed by both gene deletion and shuffling from the S

blocks, 3) duplicate deletion/movement may have been mediated by

single-/double-stranded illegitimate recombination mechanisms, 4) modern genomes

arose through centromeric fusion of protochromosomes, leading to functional

monocentric neochromosomes, 5) the fusion of two dominant blocks leads to

supradominant neochromosomes (D + D = D) with higher ancestral gene retention

compared with D + S = D (i.e., fusion of blocks with opposite sensitivity) or

even S + S = S (i.e., fusion of two sensitive ancestral blocks). A new

user-friendly online tool named "PlantSyntenyViewer," available at

http://urgi.versailles.inra.fr/synteny-cereal, presents the refined comparative

genomics data.

DOI: 10.1093/gbe/evt200

PMCID: PMC3914691

PMID: 24317974 [Indexed for MEDLINE]

925. J Alzheimers Dis. 2014;41(2):587-97. doi: 10.3233/JAD-140147.

Robust gene dysregulation in Alzheimer's disease brains.

Feng X(1), Bai Z(2), Wang J(2), Xie B(2), Sun J(2), Han G(2), Song F(2), Crack

PJ(3), Duan Y(4), Lei H(5).

Author information:

(1)CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of

Genomics, Chinese Academy of Sciences, Beijing, China. (2)CAS Key Laboratory of

Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy

of Sciences, Beijing, China University of Chinese Academy of Sciences, Beijing,

China. (3)Department of Pharmacology and Therapeutics, University of Melbourne,

Parkville, Australia. (4)UC Davis Genome Center and Department of Biomedical

Engineering, Davis, CA, USA. (5)CAS Key Laboratory of Genome Sciences and

Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing,

China UC Davis Genome Center and Department of Biomedical Engineering, Davis, CA,

USA Center of Alzheimer's disease, Beijing Institute for Brain Disorders,

Beijing, China.

The brain transcriptome of Alzheimer's disease (AD) reflects the prevailing

disease mechanism at the gene expression level. However, thousands of genes have

been reported to be dysregulated in AD brains in existing studies, and the

consistency or discrepancy among these studies has not been thoroughly examined.

Toward this end, we conducted a comprehensive survey of the brain transcriptome

datasets for AD and other neurological diseases. We first demonstrated that the

frequency of observed dysregulation in AD was highly correlated with the

reproducibility of the dysregulation. Based on this observation, we selected 100

genes with the highest frequency of dysregulation to illustrate the core

perturbation in AD brains. The dysregulation of these genes was validated in

several independent datasets for AD. We further identified 12 genes with strong

correlation of gene expression with disease progression. The relevance of these

genes to disease progression was also validated in an independent dataset.

Interestingly, we found a transcriptional "cushion" for these 100 genes in the

less vulnerable visual cortex region, which may be a critical component of the

protection mechanism for less vulnerable brain regions. To facilitate the

research in this field, we have provided the expression information of ~8000

relevant genes on a publicly accessible web server AlzBIG (http://alz.big.ac.cn).

DOI: 10.3233/JAD-140147

PMID: 24662101 [Indexed for MEDLINE]

926. J Biomol Struct Dyn. 2014;32(11):1752-8. doi: 10.1080/07391102.2013.834514. Epub

2013 Sep 13.

Zebra: a web server for bioinformatic analysis of diverse protein families.

Suplatov D(1), Kirilin E, Takhaveev V, Svedas V.

Author information:

(1)a Faculty of Bioengineering and Bioinformatics , Belozersky Institute of

Physicochemical Biology, Lomonosov Moscow State University , Vorobjev hills,

1-40, Moscow , 119991 , Russia .

During evolution of proteins from a common ancestor, one functional property can

be preserved while others can vary leading to functional diversity. A systematic

study of the corresponding adaptive mutations provides a key to one of the most

challenging problems of modern structural biology - understanding the impact of

amino acid substitutions on protein function. The subfamily-specific positions

(SSPs) are conserved within functional subfamilies but are different between them

and, therefore, seem to be responsible for functional diversity in protein

superfamilies. Consequently, a corresponding method to perform the bioinformatic

analysis of sequence and structural data has to be implemented in the common

laboratory practice to study the structure-function relationship in proteins and

develop novel protein engineering strategies. This paper describes Zebra web

server - a powerful remote platform that implements a novel bioinformatic

analysis algorithm to study diverse protein families. It is the first application

that provides specificity determinants at different levels of functional

classification, therefore addressing complex functional diversity of large

superfamilies. Statistical analysis is implemented to automatically select a set

of highly significant SSPs to be used as hotspots for directed evolution or

rational design experiments and analyzed studying the structure-function

relationship. Zebra results are provided in two ways - (1) as a single all-in-one

parsable text file and (2) as PyMol sessions with structural representation of

SSPs. Zebra web server is available at http://biokinet.belozersky.msu.ru/zebra .

DOI: 10.1080/07391102.2013.834514

PMID: 24028489 [Indexed for MEDLINE]

927. J Biomol Struct Dyn. 2014;32(1):36-51. doi: 10.1080/07391102.2012.746945. Epub

2013 Jan 9.

Comprehensively designed consensus of standalone secondary structure predictors

improves Q3 by over 3%.

Yan J(1), Marcus M, Kurgan L.

Author information:

(1)a Department of Electrical and Computer Engineering , University of Alberta ,

Edmonton , Canada .

Protein fold is defined by a spatial arrangement of three types of secondary

structures (SSs) including helices, sheets, and coils/loops. Current methods that

predict SS from sequences rely on complex machine learning-derived models and

provide the three-state accuracy (Q3) at about 82%. Further improvements in

predictive quality could be obtained with a consensus-based approach, which so

far received limited attention. We perform first-of-its-kind comprehensive design

of a SS consensus predictor (SScon), in which we consider 12 modern standalone SS

predictors and utilize Support Vector Machine (SVM) to combine their predictions.

Using a large benchmark data-set with 10 random training-test splits, we show

that a simple, voting-based consensus of carefully selected base methods improves

Q3 by 1.9% when compared to the best single predictor. Use of SVM provides

additional 1.4% improvement with the overall Q3 at 85.6% and segment overlap

(SOV3) at 83.7%, when compared to 82.3 and 80.9%, respectively, obtained by the

best individual methods. We also show strong improvements when the consensus is

based on ab-initio methods, with Q3 = 82.3% and SOV3 = 80.7% that match the

results from the best template-based approaches. Our consensus reduces the number

of significant errors where helix is confused with a strand, provides

particularly good results for short helices and strands, and gives the most

accurate estimates of the content of individual SSs in the chain. Case studies

are used to visualize the improvements offered by the consensus at the residue

level. A web-server and a standalone implementation of SScon are available at

http://biomine.ece.ualberta.ca/SSCon/ .

DOI: 10.1080/07391102.2012.746945

PMID: 23298369 [Indexed for MEDLINE]

928. Methods Mol Biol. 2014;1179:279-90. doi: 10.1007/978-1-4939-1053-3\_19.

The Mutagenesis Assistant Program.

Verma R(1), Wong TS, Schwaneberg U, Roccatano D.

Author information:

(1)Department of Chemistry, Wichita State University, Wichita, Kansas, USA.

Mutagenesis Assistant Program (MAP) is a web-based statistical tool to develop

directed evolution strategies by investigating the consequences at the amino acid

level of the mutational biases of random mutagenesis methods on any given gene.

The latest development of the program, the MAP(2.0)3D server, correlates the

generated amino acid substitution patterns of a specific random mutagenesis

method to the sequence and structural information of the target protein. The

combined information can be used to select an experimental strategy that improves

the chances of obtaining functionally efficient and/or stable enzyme variants.

Hence, the MAP(2.0)3D server facilitates the "in silico" prescreening of the

target gene by predicting the amino acid diversity generated in a random

mutagenesis library. Here, we describe the features of MAP(2.0)3D server by

analyzing, as an example, the cytochrome P450BM3 monooxygenase (CYP102A1). The

MAP(2.0)3D server is available publicly at

http://map.jacobs-university.de/map3d.html.

DOI: 10.1007/978-1-4939-1053-3\_19

PMID: 25055785 [Indexed for MEDLINE]

929. Methods Mol Biol. 2014;1137:181-97. doi: 10.1007/978-1-4939-0366-5\_13.

Predicting the structure of protein-protein complexes using the SwarmDock Web

Server.

Torchala M(1), Bates PA.

Author information:

(1)Biomolecular Modelling Laboratory, Cancer Research UK London Research

Institute, London, UK.

Protein-protein interactions drive many of the biological functions of the cell.

Any two proteins have the potential to interact; however, whether the

interactions are of biological significance is dependent on a number of

complicated factors. Thus, modelling the three-dimensional structure of

protein-protein complexes is still considered to be a complex endeavor.

Nevertheless, many experimentalists now wish to boost their knowledge of

protein-protein interactions, well beyond complexes resolved experimentally, and

for them to be able to do so it is important they are able to effectively and

confidently predict protein-protein interactions. The main aim of this chapter is

to acquaint the reader, particularly one from a non-computational background, how

to use a state-of-the-art protein docking tool. In particular, we describe here

the SwarmDock Server (SDS), a web service for the flexible modelling of

protein-protein complexes; this server is freely available at:

http://bmm.cancerresearchuk.org/~SwarmDock/. Supplementary files for Case Studies

are provided with the chapter and available at extras.springer.com.

DOI: 10.1007/978-1-4939-0366-5\_13

PMID: 24573482 [Indexed for MEDLINE]

930. Methods Mol Biol. 2014;1137:119-30. doi: 10.1007/978-1-4939-0366-5\_9.

SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function

by fold recognition and binding affinity prediction.

Yang Y(1), Zhao H, Wang J, Zhou Y.

Author information:

(1)School of Informatics, Indiana University Purdue University, Indianapolis, IN,

USA.

RNA-binding proteins (RBPs) play key roles in RNA metabolism and

post-transcriptional regulation. Computational methods have been developed

separately for prediction of RBPs and RNA-binding residues by machine-learning

techniques and prediction of protein-RNA complex structures by rigid or

semiflexible structure-to-structure docking. Here, we describe a template-based

technique called SPOT-Seq-RNA that integrates prediction of RBPs, RNA-binding

residues, and protein-RNA complex structures into a single package. This

integration is achieved by combining template-based structure-prediction

software, SPARKS X, with binding affinity prediction software, DRNA. This tool

yields reasonable sensitivity (46 %) and high precision (84 %) for an independent

test set of 215 RBPs and 5,766 non-RBPs. SPOT-Seq-RNA is computationally

efficient for genome-scale prediction of RBPs and protein-RNA complex structures.

Its application to human genome study has revealed a similar sensitivity and

ability to uncover hundreds of novel RBPs beyond simple homology. The online

server and downloadable version of SPOT-Seq-RNA are available at

http://sparks-lab.org/server/SPOT-Seq-RNA/.

DOI: 10.1007/978-1-4939-0366-5\_9

PMCID: PMC3937850

PMID: 24573478 [Indexed for MEDLINE]

931. Methods Mol Biol. 2014;1137:105-17. doi: 10.1007/978-1-4939-0366-5\_8.

3D-SURFER 2.0: web platform for real-time search and characterization of protein

surfaces.

Xiong Y(1), Esquivel-Rodriguez J, Sael L, Kihara D.

Author information:

(1)Department of Biological Sciences, Purdue University, West Lafayette, IN, USA.

The increasing number of uncharacterized protein structures necessitates the

development of computational approaches for function annotation using the protein

tertiary structures. Protein structure database search is the basis of any

structure-based functional elucidation of proteins. 3D-SURFER is a web platform

for real-time protein surface comparison of a given protein structure against the

entire PDB using 3D Zernike descriptors. It can smoothly navigate the protein

structure space in real-time from one query structure to another. A major new

feature of Release 2.0 is the ability to compare the protein surface of a single

chain, a single domain, or a single complex against databases of protein chains,

domains, complexes, or a combination of all three in the latest PDB.

Additionally, two types of protein structures can now be compared:

all-atom-surface and backbone-atom-surface. The server can also accept a batch

job for a large number of database searches. Pockets in protein surfaces can be

identified by VisGrid and LIGSITE (csc) . The server is available at

http://kiharalab.org/3d-surfer/.

DOI: 10.1007/978-1-4939-0366-5\_8

PMID: 24573477 [Indexed for MEDLINE]

932. Methods Mol Biol. 2014;1137:29-41. doi: 10.1007/978-1-4939-0366-5\_3.

The MULTICOM protein tertiary structure prediction system.

Li J(1), Bhattacharya D, Cao R, Adhikari B, Deng X, Eickholt J, Cheng J.

Author information:

(1)Computer Science Department, C. Bond Life Science Center, Informatics

Institute, University of Missouri, Columbia, MO, USA.

With the expansion of genomics and proteomics data aided by the rapid progress of

next-generation sequencing technologies, computational prediction of protein

three-dimensional structure is an essential part of modern structural genomics

initiatives. Prediction of protein structure through understanding of the

theories behind protein sequence-structure relationship, however, remains one of

the most challenging problems in contemporary life sciences. Here, we describe

MULTICOM, a multi-level combination technique, intended to predict moderate- to

high-resolution structure of a protein through a novel approach of combining

multiple sources of complementary information derived from the experimentally

solved protein structures in the Protein Data Bank. The MULTICOM web server is

freely available at http://sysbio.rnet.missouri.edu/multicom\_toolbox/.

DOI: 10.1007/978-1-4939-0366-5\_3

PMID: 24573472 [Indexed for MEDLINE]

933. Methods Mol Biol. 2014;1137:17-27. doi: 10.1007/978-1-4939-0366-5\_2.

RaptorX server: a resource for template-based protein structure modeling.

Källberg M(1), Margaryan G, Wang S, Ma J, Xu J.

Author information:

(1)Toyota Technological Institute, Chicago, IL, USA.

Assigning functional properties to a newly discovered protein is a key challenge

in modern biology. To this end, computational modeling of the three-dimensional

atomic arrangement of the amino acid chain is often crucial in determining the

role of the protein in biological processes. We present a community-wide

web-based protocol, RaptorX server ( http://raptorx.uchicago.edu ), for automated

protein secondary structure prediction, template-based tertiary structure

modeling, and probabilistic alignment sampling.Given a target sequence, RaptorX

server is able to detect even remotely related template sequences by means of a

novel nonlinear context-specific alignment potential and probabilistic

consistency algorithm. Using the protocol presented here it is thus possible to

obtain high-quality structural models for many target protein sequences when only

distantly related protein domains have experimentally solved structures. At

present, RaptorX server can perform secondary and tertiary structure prediction

of a 200 amino acid target sequence in approximately 30 min.

DOI: 10.1007/978-1-4939-0366-5\_2

PMID: 24573471 [Indexed for MEDLINE]

934. Methods Mol Biol. 2014;1079:263-71. doi: 10.1007/978-1-62703-646-7\_17.

PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and

three-dimensional structural information.

Pei J(1), Grishin NV.

Author information:

(1)Howard Hughes Medical Institute, University of Texas Southwestern Medical

Center, Dallas, TX, USA.

Multiple sequence alignment (MSA) is an essential tool with many applications in

bioinformatics and computational biology. Accurate MSA construction for divergent

proteins remains a difficult computational task. The constantly increasing

protein sequences and structures in public databases could be used to improve

alignment quality. PROMALS3D is a tool for protein MSA construction enhanced with

additional evolutionary and structural information from database searches.

PROMALS3D automatically identifies homologs from sequence and structure databases

for input proteins, derives structure-based constraints from alignments of

three-dimensional structures, and combines them with sequence-based constraints

of profile-profile alignments in a consistency-based framework to construct

high-quality multiple sequence alignments. PROMALS3D output is a consensus

alignment enriched with sequence and structural information about input proteins

and their homologs. PROMALS3D Web server and package are available at

http://prodata.swmed.edu/PROMALS3D.

DOI: 10.1007/978-1-62703-646-7\_17

PMCID: PMC4506754

PMID: 24170408 [Indexed for MEDLINE]

935. Methods Mol Biol. 2014;1079:245-62. doi: 10.1007/978-1-62703-646-7\_16.

PRALINE: a versatile multiple sequence alignment toolkit.

Bawono P(1), Heringa J.

Author information:

(1)Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam,

Amsterdam, The Netherlands.

Profile ALIgNmEnt (PRALINE) is a versatile multiple sequence alignment toolkit.

In its main alignment protocol, PRALINE follows the global progressive alignment

algorithm. It provides various alignment optimization strategies to address the

different situations that call for protein multiple sequence alignment: global

profile preprocessing, homology-extended alignment, secondary structure-guided

alignment, and transmembrane aware alignment. A number of combinations of these

strategies are enabled as well. PRALINE is accessible via the online server

http://www.ibi.vu.nl/programs/PRALINEwww/. The server facilitates extensive

visualization possibilities aiding the interpretation of alignments generated,

which can be written out in pdf format for publication purposes. PRALINE also

allows the sequences in the alignment to be represented in a dendrogram to show

their mutual relationships according to the alignment. The chapter ends with a

discussion of various issues occurring in multiple sequence alignment.

DOI: 10.1007/978-1-62703-646-7\_16

PMID: 24170407 [Indexed for MEDLINE]

936. Methods Mol Biol. 2014;1079:191-202. doi: 10.1007/978-1-62703-646-7\_12.

Multiple sequence alignment with DIALIGN.

Morgenstern B(1).

Author information:

(1)Abteilung für Bioinformatik (IMG), Universitat Göttingen, Göttingen, Germany.

DIALIGN is a software tool for multiple sequence alignment by combining global

and local alignment features. It composes multiple alignments from local pairwise

sequence similarities. This approach is particularly useful to discover conserved

functional regions in sequences that share only local homologies but are

otherwise unrelated. An anchoring option allows to use external information and

expert knowledge in addition to primary-sequence similarity alone. The latest

version of DIALIGN optionally uses matches to the PFAM database to detect weak

homologies. Various versions of the program are available through Göttingen

Bioinformatics Compute Server (GOBICS) at

http://www.gobics.de/department/software.

DOI: 10.1007/978-1-62703-646-7\_12

PMID: 24170403 [Indexed for MEDLINE]

937. Methods Mol Biol. 2014;1079:171-89. doi: 10.1007/978-1-62703-646-7\_11.

GramAlign: fast alignment driven by grammar-based phylogeny.

Russell DJ(1).

Author information:

(1)Department of Electrical Engineering, University of Nebraska-Lincoln, Lincoln,

NE, USA.

Multiple sequence alignment involves identifying related subsequences among

biological sequences. When matches are found, the associated pieces are shifted

so that when sequences are presented as successive rows-one sequence per

row-homologous residues line-up in columns. Exact alignment of more than a few

sequences is known to be computationally prohibitive. Thus many heuristic

algorithms have been developed to produce good alignments in an efficient amount

of time by determining an order by which pairs of sequences are progressively

aligned and merged. GRAMALIGN is such a progressive alignment algorithm that uses

a grammar-based relative complexity distance metric to determine the alignment

order. This technique allows for a computationally efficient and scalable program

useful for aligning both large numbers of sequences and sets of long sequences

quickly. The GRAMALIGN software is available at

http://bioinfo.unl.edu/gramalign.php for both source code download and a

web-based alignment server.

DOI: 10.1007/978-1-62703-646-7\_11

PMID: 24170402 [Indexed for MEDLINE]

938. Methods Mol Biol. 2014;1079:117-29. doi: 10.1007/978-1-62703-646-7\_7.

T-Coffee: Tree-based consistency objective function for alignment evaluation.

Magis C(1), Taly JF, Bussotti G, Chang JM, Di Tommaso P, Erb I, Espinosa-Carrasco

J, Notredame C.

Author information:

(1)Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG),

Universitat Pompeu Fabra, Barcelona, Spain.

T-Coffee, for Tree-based consistency objective function for alignment evaluation,

is a versatile multiple sequence alignment (MSA) method suitable for aligning

virtually any type of biological sequences. T-Coffee provides more than a simple

sequence aligner; rather it is a framework in which alternative alignment methods

and/or extra information (i.e., structural, evolutionary, or experimental

information) can be combined to reach more accurate and more meaningful MSAs.

T-Coffee can be used either by running input data via the Web server (

http://tcoffee.crg.cat/apps/tcoffee/index.html ) or by downloading the T-Coffee

package. Here, we present how the package can be used in its command line mode to

carry out the most common tasks and multiply align proteins, DNA, and RNA

sequences. This chapter particularly emphasizes on the description of T-Coffee

special flavors also called "modes," designed to address particular biological

problems.

DOI: 10.1007/978-1-62703-646-7\_7

PMID: 24170398 [Indexed for MEDLINE]

939. Methods Mol Biol. 2014;1084:159-72. doi: 10.1007/978-1-62703-658-0\_9.

Analysis of protein conformational transitions using elastic network model.

Zheng W(1), Tekpinar M.

Author information:

(1)Department of Physics, University at Buffalo, Buffalo, NY, USA.

In this chapter, we demonstrate the usage of a coarse-grained elastic network

model to analyze protein conformational transitions in the NS3 helicase (NS3hel)

of Hepatitis C virus (HCV). This analysis allows us to identify and visualize

collective domain motions involved in the conformational transitions and predict

the order of structural events during the transitions. It is highly efficient and

applicable to many multi-domain protein structures which undergo large

conformational changes to fulfill their functions. This method is made available

through a Web server ( http://enm.lobos.nih.gov ).

DOI: 10.1007/978-1-62703-658-0\_9

PMID: 24061921 [Indexed for MEDLINE]

940. Nat Protoc. 2014 Jan;9(1):156-70. doi: 10.1038/nprot.2013.172. Epub 2013 Dec 19.

Validation of metal-binding sites in macromolecular structures with the

CheckMyMetal web server.

Zheng H(1), Chordia MD(1), Cooper DR(1), Chruszcz M(2), Müller P(3), Sheldrick

GM(4), Minor W(1).

Author information:

(1)1] Department of Molecular Physiology and Biological Physics, University of

Virginia, Charlottesville, Virginia, USA. [2] Center for Structural Genomics of

Infectious Diseases (CSGID) Consortium, USA. (2)1] Department of Molecular

Physiology and Biological Physics, University of Virginia, Charlottesville,

Virginia, USA. [2] Center for Structural Genomics of Infectious Diseases (CSGID)

Consortium, USA. [3] Department of Chemistry and Biochemistry, University of

South Carolina, Columbia, South Carolina, USA. (3)Department of Chemistry,

Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

(4)Lehrstuhl für Strukturchemie, Universität Göttingen, Göttingen, Germany.

Metals have vital roles in both the mechanism and architecture of biological

macromolecules. Yet structures of metal-containing macromolecules in which metals

are misidentified and/or suboptimally modeled are abundant in the Protein Data

Bank (PDB). This shows the need for a diagnostic tool to identify and correct

such modeling problems with metal-binding environments. The CheckMyMetal (CMM)

web server (http://csgid.org/csgid/metal\_sites/) is a sophisticated,

user-friendly web-based method to evaluate metal-binding sites in macromolecular

structures using parameters derived from 7,350 metal-binding sites observed in a

benchmark data set of 2,304 high-resolution crystal structures. The protocol

outlines how the CMM server can be used to detect geometric and other

irregularities in the structures of metal-binding sites, as well as how it can

alert researchers to potential errors in metal assignment. The protocol also

gives practical guidelines for correcting problematic sites by modifying the

metal-binding environment and/or redefining metal identity in the PDB file.

Several examples where this has led to meaningful results are described in the

ANTICIPATED RESULTS section. CMM was designed for a broad audience--biomedical

researchers studying metal-containing proteins and nucleic acids--but it is

equally well suited for structural biologists validating new structures during

modeling or refinement. The CMM server takes the coordinates of a

metal-containing macromolecule structure in the PDB format as input and responds

within a few seconds for a typical protein structure with 2-5 metal sites and a

few hundred amino acids.

DOI: 10.1038/nprot.2013.172

PMCID: PMC4410975

PMID: 24356774 [Indexed for MEDLINE]

941. Nucleic Acids Res. 2014 Jan;42(Database issue):D206-14. doi: 10.1093/nar/gkt1226.

Epub 2013 Nov 29.

The SEED and the Rapid Annotation of microbial genomes using Subsystems

Technology (RAST).

Overbeek R(1), Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes

S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R.

Author information:

(1)Fellowship for Interpretation of Genomes, Burr Ridge, IL 60527, USA,

Mathematics and Computer Science Division, Argonne National Laboratory, Argonne,

IL 60439, USA, Computation Institute, University of Chicago, Chicago, IL 60637,

USA, Department of Microbiology, University of Illinois at Urbana-Champaign,

Urbana, IL 61801, USA, Department of Computer Science, San Diego State

University, San Diego, CA 92182, USA, Virginia Bioinformatics Institute, Virginia

Tech, Blacksburg, VA 24060, USA, Computing, Environment and Life Sciences,

Argonne National Laboratory, Argonne, IL 60439, USA and Department of Computer

Science, University of Chicago, Chicago, IL 60637, USA.

In 2004, the SEED (http://pubseed.theseed.org/) was created to provide consistent

and accurate genome annotations across thousands of genomes and as a platform for

discovering and developing de novo annotations. The SEED is a constantly updated

integration of genomic data with a genome database, web front end, API and server

scripts. It is used by many scientists for predicting gene functions and

discovering new pathways. In addition to being a powerful database for

bioinformatics research, the SEED also houses subsystems (collections of

functionally related protein families) and their derived FIGfams (protein

families), which represent the core of the RAST annotation engine

(http://rast.nmpdr.org/). When a new genome is submitted to RAST, genes are

called and their annotations are made by comparison to the FIGfam collection. If

the genome is made public, it is then housed within the SEED and its proteins

populate the FIGfam collection. This annotation cycle has proven to be a robust

and scalable solution to the problem of annotating the exponentially increasing

number of genomes. To date, >12 000 users worldwide have annotated >60 000

distinct genomes using RAST. Here we describe the interconnectedness of the SEED

database and RAST, the RAST annotation pipeline and updates to both resources.

DOI: 10.1093/nar/gkt1226

PMCID: PMC3965101

PMID: 24293654 [Indexed for MEDLINE]

942. Nucleic Acids Res. 2014 Jan;42(Database issue):D336-46. doi: 10.1093/nar/gkt1144.

Epub 2013 Nov 23.

ModBase, a database of annotated comparative protein structure models and

associated resources.

Pieper U(1), Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, Khuri N,

Spill YG, Weinkam P, Hammel M, Tainer JA, Nilges M, Sali A.

Author information:

(1)Department of Bioengineering and Therapeutic Sciences, California Institute

for Quantitative Biosciences, Byers Hall at Mission Bay, Office 503B, University

of California at San Francisco, 1700 4th Street, San Francisco, CA 94158, USA,

Department of Pharmaceutical Chemistry, California Institute for Quantitative

Biosciences, Byers Hall at Mission Bay, Office 503B, University of California at

San Francisco, 1700 4th Street, San Francisco, CA 94158, USA, Graduate Group in

Biophysics, University of California at San Francisco, CA 94158, USA, Structural

Bioinformatics Unit, Structural Biology and Chemistry department, Institut

Pasteur, 25 rue du Docteur Roux, 75015 Paris, France, Université Paris

Diderot-Paris 7, école doctorale iViv, Paris Rive Gauche, 5 rue Thomas Mann,

75013 Paris, France, Physical Biosciences Division, Lawrence Berkeley National

Laboratory, Berkeley, CA 94720, USA, Department of Molecular Biology, Skaggs

Institute of Chemical Biology, The Scripps Research Institute, La Jolla, CA

92037, USA, Life Sciences Division, Department of Molecular Biology, Lawrence

Berkeley National Laboratory, Berkeley, CA 94720, USA.

ModBase (http://salilab.org/modbase) is a database of annotated comparative

protein structure models. The models are calculated by ModPipe, an automated

modeling pipeline that relies primarily on Modeller for fold assignment,

sequence-structure alignment, model building and model assessment

(http://salilab.org/modeller/). ModBase currently contains almost 30 million

reliable models for domains in 4.7 million unique protein sequences. ModBase

allows users to compute or update comparative models on demand, through an

interface to the ModWeb modeling server (http://salilab.org/modweb). ModBase

models are also available through the Protein Model Portal

(http://www.proteinmodelportal.org/). Recently developed associated resources

include the AllosMod server for modeling ligand-induced protein dynamics

(http://salilab.org/allosmod), the AllosMod-FoXS server for predicting a

structural ensemble that fits an SAXS profile (http://salilab.org/allosmod-foxs),

the FoXSDock server for protein-protein docking filtered by an SAXS profile

(http://salilab.org/foxsdock), the SAXS Merge server for automatic merging of

SAXS profiles (http://salilab.org/saxsmerge) and the Pose & Rank server for

scoring protein-ligand complexes (http://salilab.org/poseandrank). In this

update, we also highlight two applications of ModBase: a PSI:Biology initiative

to maximize the structural coverage of the human alpha-helical transmembrane

proteome and a determination of structural determinants of human immunodeficiency

virus-1 protease specificity.

DOI: 10.1093/nar/gkt1144

PMCID: PMC3965011

PMID: 24271400 [Indexed for MEDLINE]

943. Nucleic Acids Res. 2014 Jan;42(Database issue):D167-71. doi: 10.1093/nar/gkt1165.

Epub 2013 Nov 22.

OnTheFly: a database of Drosophila melanogaster transcription factors and their

binding sites.

Shazman S(1), Lee H, Socol Y, Mann RS, Honig B.

Author information:

(1)Howard Hughes Medical Institute, Department of Biochemistry and Molecular

Biophysics, Department of Systems Biology, Center for Computational Biology and

Bioinformatics, Columbia University, 1130 St. Nicholas Avenue, New York, NY

10032, USA, Department of Life Science, Open University of Israel, Ra'anana

43107, Israel and Department of Biochemistry and Molecular Biophysics, Columbia

University, 701 West 168th Street, HHSC 1104, New York, NY 10032, USA.

We present OnTheFly (http://bhapp.c2b2.columbia.edu/OnTheFly/index.php), a

database comprising a systematic collection of transcription factors (TFs) of

Drosophila melanogaster and their DNA-binding sites. TFs predicted in the

Drosophila melanogaster genome are annotated and classified and their structures,

obtained via experiment or homology models, are provided. All known preferred TF

DNA-binding sites obtained from the B1H, DNase I and SELEX methodologies are

presented. DNA shape parameters predicted for these sites are obtained from a

high throughput server or from crystal structures of protein-DNA complexes where

available. An important feature of the database is that all DNA-binding domains

and their binding sites are fully annotated in a eukaryote using structural

criteria and evolutionary homology. OnTheFly thus provides a comprehensive view

of TFs and their binding sites that will be a valuable resource for deciphering

non-coding regulatory DNA.

DOI: 10.1093/nar/gkt1165

PMCID: PMC3965123

PMID: 24271386 [Indexed for MEDLINE]

944. Nucleic Acids Res. 2014 Jan;42(Database issue):D53-9. doi: 10.1093/nar/gkt1202.

Epub 2013 Nov 22.

NGSmethDB: an updated genome resource for high quality, single-cytosine

resolution methylomes.

Geisen S(1), Barturen G, Alganza ÁM, Hackenberg M, Oliver JL.

Author information:

(1)Facultad de Ciencias, Departmento de Genética, Universidad de Granada,

18071-Granada, Spain and Laboratorio de Bioinformática, Instituto de

Biotecnología, Centro de Investigación Biomédica, 18100-Granada, Spain.

The updated release of 'NGSmethDB' (http://bioinfo2.ugr.es/NGSmethDB) is a

repository for single-base whole-genome methylome maps for the best-assembled

eukaryotic genomes. Short-read data sets from NGS bisulfite-sequencing projects

of cell lines, fresh and pathological tissues are first pre-processed and aligned

to the corresponding reference genome, and then the cytosine methylation levels

are profiled. One major improvement is the application of a unique bioinformatics

protocol to all data sets, thereby assuring the comparability of all values with

each other. We implemented stringent quality controls to minimize important error

sources, such as sequencing errors, bisulfite failures, clonal reads or single

nucleotide variants (SNVs). This leads to reliable and high-quality methylomes,

all obtained under uniform settings. Another significant improvement is the

detection in parallel of SNVs, which might be crucial for many downstream

analyses (e.g. SNVs and differential-methylation relationships). A

next-generation methylation browser allows fast and smooth scrolling and zooming,

thus speeding data download/upload, at the same time requiring fewer server

resources. Several data mining tools allow the comparison/retrieval of

methylation levels in different tissues or genome regions. NGSmethDB methylomes

are also available as native tracks through a UCSC hub, which allows comparison

with a wide range of third-party annotations, in particular phenotype or disease

annotations.

DOI: 10.1093/nar/gkt1202

PMCID: PMC3964946

PMID: 24271385 [Indexed for MEDLINE]

945. Nucleic Acids Res. 2014 Jan;42(Database issue):D780-8. doi: 10.1093/nar/gkt1092.

Epub 2013 Nov 13.

FlyBase 102--advanced approaches to interrogating FlyBase.

St Pierre SE(1), Ponting L, Stefancsik R, McQuilton P; FlyBase Consortium.

Author information:

(1)The Biological Laboratories, Harvard University, 16 Divinity Avenue,

Cambridge, MA 02138, USA and Department of Genetics, University of Cambridge,

Downing Street, Cambridge CB2 3EH, UK.

FlyBase (http://flybase.org) is the leading website and database of Drosophila

genes and genomes. Whether you are using the fruit fly Drosophila melanogaster as

an experimental system or wish to understand Drosophila biological knowledge in

relation to human disease or to other model systems, FlyBase can help you

successfully find the information you are looking for. Here, we demonstrate some

of our more advanced searching systems and highlight some of our new tools for

searching the wealth of data on FlyBase. The first section explores gene function

in FlyBase, using our TermLink tool to search with Controlled Vocabulary terms

and our new RNA-Seq Search tool to search gene expression. The second section of

this article describes a few ways to search genomic data in FlyBase, using our

BLAST server and the new implementation of GBrowse 2, as well as our new

FeatureMapper tool. Finally, we move on to discuss our most powerful search tool,

QueryBuilder, before describing pre-computed cuts of the data and how to query

the database programmatically.

DOI: 10.1093/nar/gkt1092

PMCID: PMC3964969

PMID: 24234449 [Indexed for MEDLINE]

946. Nucleic Acids Res. 2014 Jan;42(Database issue):D1182-7. doi: 10.1093/nar/gkt1016.

Epub 2013 Oct 29.

PlantTFDB 3.0: a portal for the functional and evolutionary study of plant

transcription factors.

Jin J(1), Zhang H, Kong L, Gao G, Luo J.

Author information:

(1)State Key Laboratory of Protein and Plant Gene Research, College of Life

Sciences and Center for Bioinformatics, Peking University, Beijing 100871, P.R.

China.

With the aim to provide a resource for functional and evolutionary study of plant

transcription factors (TFs), we updated the plant TF database PlantTFDB to

version 3.0 (http://planttfdb.cbi.pku.edu.cn). After refining the TF

classification pipeline, we systematically identified 129 288 TFs from 83

species, of which 67 species have genome sequences, covering main lineages of

green plants. Besides the abundant annotation provided in the previous version,

we generated more annotations for identified TFs, including expression,

regulation, interaction, conserved elements, phenotype information,

expert-curated descriptions derived from UniProt, TAIR and NCBI GeneRIF, as well

as references to provide clues for functional studies of TFs. To help identify

evolutionary relationship among identified TFs, we assigned 69 450 TFs into 3924

orthologous groups, and constructed 9217 phylogenetic trees for TFs within the

same families or same orthologous groups, respectively. In addition, we set up a

TF prediction server in this version for users to identify TFs from their own

sequences.

DOI: 10.1093/nar/gkt1016

PMCID: PMC3965000

PMID: 24174544 [Indexed for MEDLINE]

947. Nucleic Acids Res. 2014 Jan;42(Database issue):D273-8. doi: 10.1093/nar/gkt927.

Epub 2013 Oct 22.

HRaP: database of occurrence of HomoRepeats and patterns in proteomes.

Lobanov MY(1), Sokolovskiy IV, Galzitskaya OV.

Author information:

(1)Group of Bioinformatics, Institute of Protein Research, Russian Academy of

Sciences, Pushchino, Moscow Region 142290, Russia.

We focus our attention on multiple repeats of one amino acid (homorepeats) and

create a new database (named HRaP, at http://bioinfo.protres.ru/hrap/) of

occurrence of homorepeats and disordered patterns in different proteomes. HRaP is

aimed at understanding the amino acid tandem repeat function in different

proteomes. Therefore, the database includes 122 proteomes, 97 eukaryotic and 25

bacterial ones that can be divided into 9 kingdoms and 5 phyla of bacteria. The

database includes 1,449,561 protein sequences and 771,786 sequences of proteins

with GO annotations. We have determined homorepeats and patterns that are

associated with some function. Through our web server, the user can do the

following: (i) search for proteins with the given homorepeat in 122 proteomes,

including GO annotation for these proteins; (ii) search for proteins with the

given disordered pattern from the library of disordered patterns constructed on

the clustered Protein Data Bank in 122 proteomes, including GO annotations for

these proteins; (iii) analyze lengths of homorepeats in different proteomes; (iv)

investigate disordered regions in the chosen proteins in 122 proteomes; (v) study

the coupling of different homorepeats in one protein; (vi) determine longest runs

for each amino acid inside each proteome; and (vii) download the full list of

proteins with the given length of a homorepeat.

DOI: 10.1093/nar/gkt927

PMCID: PMC3965023

PMID: 24150944 [Indexed for MEDLINE]

948. Nucleic Acids Res. 2014 Jan;42(1):97-108. doi: 10.1093/nar/gkt890. Epub 2013 Oct

3.

De novo prediction of DNA-binding specificities for Cys2His2 zinc finger

proteins.

Persikov AV(1), Singh M.

Author information:

(1)Lewis-Sigler Institute for Integrative Genomics, Princeton University,

Princeton NJ 08544, USA and Department of Computer Science, Princeton University,

Princeton NJ 08544, USA.

Proteins with sequence-specific DNA binding function are important for a wide

range of biological activities. De novo prediction of their DNA-binding

specificities from sequence alone would be a great aid in inferring cellular

networks. Here we introduce a method for predicting DNA-binding specificities for

Cys2His2 zinc fingers (C2H2-ZFs), the largest family of DNA-binding proteins in

metazoans. We develop a general approach, based on empirical calculations of

pairwise amino acid-nucleotide interaction energies, for predicting position

weight matrices (PWMs) representing DNA-binding specificities for C2H2-ZF

proteins. We predict DNA-binding specificities on a per-finger basis and merge

predictions for C2H2-ZF domains that are arrayed within sequences. We test our

approach on a diverse set of natural C2H2-ZF proteins with known binding

specificities and demonstrate that for >85% of the proteins, their predicted PWMs

are accurate in 50% of their nucleotide positions. For proteins with several zinc

finger isoforms, we show via case studies that this level of accuracy enables us

to match isoforms with their known DNA-binding specificities. A web server for

predicting a PWM given a protein containing C2H2-ZF domains is available online

at http://zf.princeton.edu and can be used to aid in protein engineering

applications and in genome-wide searches for transcription factor targets.

DOI: 10.1093/nar/gkt890

PMCID: PMC3874201

PMID: 24097433 [Indexed for MEDLINE]

949. PLoS Comput Biol. 2014 Jan;10(1):e1003440. doi: 10.1371/journal.pcbi.1003440.

Epub 2014 Jan 16.

PredictSNP: robust and accurate consensus classifier for prediction of

disease-related mutations.

Bendl J(1), Stourac J(2), Salanda O(3), Pavelka A(4), Wieben ED(5), Zendulka

J(3), Brezovsky J(4), Damborsky J(2).

Author information:

(1)Loschmidt Laboratories, Department of Experimental Biology and Research Centre

for Toxic Compounds in the Environment, Faculty of Science, Masaryk University,

Brno, Czech Republic ; Department of Information Systems, Faculty of Information

Technology, Brno University of Technology, Brno, Czech Republic ; Center of

Biomolecular and Cellular Engineering, International Centre for Clinical

Research, St. Anne's University Hospital Brno, Brno, Czech Republic. (2)Loschmidt

Laboratories, Department of Experimental Biology and Research Centre for Toxic

Compounds in the Environment, Faculty of Science, Masaryk University, Brno, Czech

Republic ; Center of Biomolecular and Cellular Engineering, International Centre

for Clinical Research, St. Anne's University Hospital Brno, Brno, Czech Republic.

(3)Department of Information Systems, Faculty of Information Technology, Brno

University of Technology, Brno, Czech Republic. (4)Loschmidt Laboratories,

Department of Experimental Biology and Research Centre for Toxic Compounds in the

Environment, Faculty of Science, Masaryk University, Brno, Czech Republic.

(5)Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, New

York, United States of America.

Single nucleotide variants represent a prevalent form of genetic variation.

Mutations in the coding regions are frequently associated with the development of

various genetic diseases. Computational tools for the prediction of the effects

of mutations on protein function are very important for analysis of single

nucleotide variants and their prioritization for experimental characterization.

Many computational tools are already widely employed for this purpose.

Unfortunately, their comparison and further improvement is hindered by large

overlaps between the training datasets and benchmark datasets, which lead to

biased and overly optimistic reported performances. In this study, we have

constructed three independent datasets by removing all duplicities,

inconsistencies and mutations previously used in the training of evaluated tools.

The benchmark dataset containing over 43,000 mutations was employed for the

unbiased evaluation of eight established prediction tools: MAPP, nsSNPAnalyzer,

PANTHER, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT and SNAP. The six best performing

tools were combined into a consensus classifier PredictSNP, resulting into

significantly improved prediction performance, and at the same time returned

results for all mutations, confirming that consensus prediction represents an

accurate and robust alternative to the predictions delivered by individual tools.

A user-friendly web interface enables easy access to all eight prediction tools,

the consensus classifier PredictSNP and annotations from the Protein Mutant

Database and the UniProt database. The web server and the datasets are freely

available to the academic community at http://loschmidt.chemi.muni.cz/predictsnp.

DOI: 10.1371/journal.pcbi.1003440

PMCID: PMC3894168

PMID: 24453961 [Indexed for MEDLINE]

950. Protein Pept Lett. 2014;21(8):736-42.

SVM-PB-Pred: SVM based protein block prediction method using sequence profiles

and secondary structures.

Suresh V, Parthasarathy S(1).

Author information:

(1)Department of Bioinformatics, School of Life Sciences, Bharathidasan

University, Tiruchirappalli 620 024, Tamil Nadu, India. bdupartha@gmail.com.

We developed a support vector machine based web server called SVM-PB-Pred, to

predict the Protein Block for any given amino acid sequence. The input features

of SVM-PB-Pred include i) sequence profiles (PSSM) and ii) actual secondary

structures (SS) from DSSP method or predicted secondary structures from NPS@ and

GOR4 methods. There were three combined input features PSSM+SS(DSSP),

PSSM+SS(NPS@) and PSSM+SS(GOR4) used to test and train the SVM models. Similarly,

four datasets RS90, DB433, LI1264 and SP1577 were used to develop the SVM models.

These four SVM models developed were tested using three different benchmarking

tests namely; (i) self consistency, (ii) seven fold cross validation test and

(iii) independent case test. The maximum possible prediction accuracy of ~70% was

observed in self consistency test for the SVM models of both LI1264 and SP1577

datasets, where PSSM+SS(DSSP) input features was used to test. The prediction

accuracies were reduced to ~53% for PSSM+SS(NPS@) and ~43% for PSSM+SS(GOR4) in

independent case test, for the SVM models of above two same datasets. Using our

method, it is possible to predict the protein block letters for any query protein

sequence with ~53% accuracy, when the SP1577 dataset and predicted secondary

structure from NPS@ server were used. The SVM-PB-Pred server can be freely

accessed through http://bioinfo.bdu.ac.in/~svmpbpred.

PMID: 23855661 [Indexed for MEDLINE]

951. RNA Biol. 2014;11(12):1619-29. doi: 10.4161/15476286.2014.992273.

Distribution and frequencies of post-transcriptional modifications in tRNAs.

Machnicka MA(1), Olchowik A, Grosjean H, Bujnicki JM.

Author information:

(1)a International Institute of Molecular and Cell Biology ; Warsaw , Poland.

Functional tRNA molecules always contain a wide variety of post-transcriptionally

modified nucleosides. These modifications stabilize tRNA structure, allow for

proper interaction with other macromolecules and fine-tune the decoding of mRNAs

during translation. Their presence in functionally important regions of tRNA is

conserved in all domains of life. However, the identities of many of these

modified residues depend much on the phylogeny of organisms the tRNAs are found

in, attesting for domain-specific strategies of tRNA maturation. In this work we

present a new tool, tRNAmodviz web server (http://genesilico.pl/trnamodviz) for

easy comparative analysis and visualization of modification patterns in

individual tRNAs, as well as in groups of selected tRNA sequences. We also

present results of comparative analysis of tRNA sequences derived from 7

phylogenetically distinct groups of organisms: Gram-negative bacteria,

Gram-positive bacteria, cytosol of eukaryotic single cell organisms, Fungi and

Metazoa, cytosol of Viridiplantae, mitochondria, plastids and Euryarchaeota.

These data update the study conducted 20 y ago with the tRNA sequences available

at that time.

DOI: 10.4161/15476286.2014.992273

PMCID: PMC4615829

PMID: 25611331 [Indexed for MEDLINE]

952. RNA Biol. 2014;11(6):693-701. Epub 2014 Apr 25.

MicroRNA binding sites in C. elegans 3' UTRs.

Liu C(1), Rennie WA(1), Mallick B(1), Kanoria S(1), Long D(1), Wolenc A(1),

Carmack CS(1), Ding Y(1).

Author information:

(1)Wadsworth Center; New York State Department of Health; Center for Medical

Science; Albany, NY USA.

MicroRNAs (miRNAs) are post-transcriptional regulators of gene expression. Since

the discovery of lin-4, the founding member of the miRNA family, over 360 miRNAs

have been identified for Caenorhabditis elegans (C. elegans). Prediction and

validation of targets are essential for elucidation of regulatory functions of

these miRNAs. For C. elegans, crosslinking immunoprecipitation (CLIP) has been

successfully performed for the identification of target mRNA sequences bound by

Argonaute protein ALG-1. In addition, reliable annotation of the 3' untranslated

regions (3' UTRs) as well as developmental stage-specific expression profiles for

both miRNAs and 3' UTR isoforms are available. By utilizing these data, we

developed statistical models and bioinformatics tools for both

transcriptome-scale and developmental stage-specific predictions of miRNA binding

sites in C. elegans 3' UTRs. In performance evaluation via cross validation on

the ALG-1 CLIP data, the models were found to offer major improvements over

established algorithms for predicting both seed sites and seedless sites. In

particular, our top-ranked predictions have a substantially higher true positive

rate, suggesting a much higher likelihood of positive experimental validation. A

gene ontology analysis of stage-specific predictions suggests that miRNAs are

involved in dynamic regulation of biological functions during C. elegans

development. In particular, miRNAs preferentially target genes related to

development, cell cycle, trafficking, and cell signaling processes. A database

for both transcriptome-scale and stage-specific predictions and software for

implementing the prediction models are available through the Sfold web server at

http://sfold.wadsworth.org.

PMCID: PMC4156501

PMID: 24827614 [Indexed for MEDLINE]

953. J Immunol Methods. 2013 Dec 31;400-401:37-44. doi: 10.1016/j.jim.2013.08.014.

Epub 2013 Aug 31.

BlockLogo: visualization of peptide and sequence motif conservation.

Olsen LR(1), Kudahl UJ, Simon C, Sun J, Schönbach C, Reinherz EL, Zhang GL,

Brusic V.

Author information:

(1)Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School,

Boston, MA, USA; Bioinformatics Centre, Department of Biology, University of

Copenhagen, Copenhagen, Denmark.

BlockLogo is a web-server application for the visualization of protein and

nucleotide fragments, continuous protein sequence motifs, and discontinuous

sequence motifs using calculation of block entropy from multiple sequence

alignments. The user input consists of a multiple sequence alignment, selection

of motif positions, type of sequence, and output format definition. The output

has BlockLogo along with the sequence logo, and a table of motif frequencies. We

deployed BlockLogo as an online application and have demonstrated its utility

through examples that show visualization of T-cell epitopes and B-cell epitopes

(both continuous and discontinuous). Our additional example shows a visualization

and analysis of structural motifs that determine the specificity of peptide

binding to HLA-DR molecules. The BlockLogo server also employs selected

experimentally validated prediction algorithms to enable on-the-fly prediction of

MHC binding affinity to 15 common HLA class I and class II alleles as well as

visual analysis of discontinuous epitopes from multiple sequence alignments. It

enables the visualization and analysis of structural and functional motifs that

are usually described as regular expressions. It provides a compact view of

discontinuous motifs composed of distant positions within biological sequences.

BlockLogo is available at: http://research4.dfci.harvard.edu/cvc/blocklogo/ and

http://met-hilab.bu.edu/blocklogo/.

© 2013.

DOI: 10.1016/j.jim.2013.08.014

PMCID: PMC3856553

PMID: 24001880 [Indexed for MEDLINE]

954. PLoS One. 2013 Dec 31;8(12):e83784. doi: 10.1371/journal.pone.0083784.

eCollection 2013.

Low-bandwidth and non-compute intensive remote identification of microbes from

raw sequencing reads.

Gautier L(1), Lund O(2).

Author information:

(1)Center for Biological Sequence Analysis (CBS), Department of Systems Biology,

Technical University of Denmark, Lyngby, Denmark ; DTU Multiassay Core (DMAC),

Technical University of Denmark, Lyngby, Denmark. (2)Center for Biological

Sequence Analysis (CBS), Department of Systems Biology, Technical University of

Denmark, Lyngby, Denmark.

Cheap DNA sequencing may soon become routine not only for human genomes but also

for practically anything requiring the identification of living organisms from

their DNA: tracking of infectious agents, control of food products, bioreactors,

or environmental samples. We propose a novel general approach to the analysis of

sequencing data where a reference genome does not have to be specified. Using a

distributed architecture we are able to query a remote server for hints about

what the reference might be, transferring a relatively small amount of data. Our

system consists of a server with known reference DNA indexed, and a client with

raw sequencing reads. The client sends a sample of unidentified reads, and in

return receives a list of matching references. Sequences for the references can

be retrieved and used for exhaustive computation on the reads, such as alignment.

To demonstrate this approach we have implemented a web server, indexing tens of

thousands of publicly available genomes and genomic regions from various

organisms and returning lists of matching hits from query sequencing reads. We

have also implemented two clients: one running in a web browser, and one as a

python script. Both are able to handle a large number of sequencing reads and

from portable devices (the browser-based running on a tablet), perform its task

within seconds, and consume an amount of bandwidth compatible with mobile

broadband networks. Such client-server approaches could develop in the future,

allowing a fully automated processing of sequencing data and routine instant

quality check of sequencing runs from desktop sequencers. A web access is

available at http://tapir.cbs.dtu.dk. The source code for a python command-line

client, a server, and supplementary data are available at http://bit.ly/1aURxkc.

DOI: 10.1371/journal.pone.0083784

PMCID: PMC3877093

PMID: 24391826 [Indexed for MEDLINE]

955. PLoS One. 2013 Dec 30;8(12):e82890. doi: 10.1371/journal.pone.0082890.

eCollection 2013.

SiBIC: a web server for generating gene set networks based on biclusters obtained

by maximal frequent itemset mining.

Takahashi K(1), Takigawa I(2), Mamitsuka H(1).

Author information:

(1)Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji,

Kyoto, Japan. (2)Bioinformatics Center, Institute for Chemical Research, Kyoto

University, Uji, Kyoto, Japan ; Creative Research Institution, Hokkaido

University, Sapporo, Hokkaido, Japan.

Detecting biclusters from expression data is useful, since biclusters are

coexpressed genes under only part of all given experimental conditions. We

present a software called SiBIC, which from a given expression dataset, first

exhaustively enumerates biclusters, which are then merged into rather independent

biclusters, which finally are used to generate gene set networks, in which a gene

set assigned to one node has coexpressed genes. We evaluated each step of this

procedure: 1) significance of the generated biclusters biologically and

statistically, 2) biological quality of merged biclusters, and 3) biological

significance of gene set networks. We emphasize that gene set networks, in which

nodes are not genes but gene sets, can be more compact than usual gene networks,

meaning that gene set networks are more comprehensible. SiBIC is available at

http://utrecht.kuicr.kyoto-u.ac.jp:8080/miami/faces/index.jsp.

DOI: 10.1371/journal.pone.0082890

PMCID: PMC3875427

PMID: 24386124 [Indexed for MEDLINE]

956. BMC Genomics. 2013 Dec 27;14:924. doi: 10.1186/1471-2164-14-924.

PSP: rapid identification of orthologous coding genes under positive selection

across multiple closely related prokaryotic genomes.

Su F, Ou HY, Tao F, Tang H, Xu P(1).

Author information:

(1)State Key Laboratory of Microbial Metabolism, Shanghai Jiao Tong University,

Shanghai 200240, P,R, China. pingxu@sjtu.edu.cn.

BACKGROUND: With genomic sequences of many closely related bacterial strains made

available by deep sequencing, it is now possible to investigate trends in

prokaryotic microevolution. Positive selection is a sub-process of

microevolution, in which a particular mutation is favored, causing the allele

frequency to continuously shift in one direction. Wide scanning of prokaryotic

genomes has shown that positive selection at the molecular level is much more

frequent than expected. Genes with significant positive selection may play key

roles in bacterial adaption to different environmental pressures. However,

selection pressure analyses are computationally intensive and awkward to

configure.

RESULTS: Here we describe an open access web server, which is designated as PSP

(Positive Selection analysis for Prokaryotic genomes) for performing evolutionary

analysis on orthologous coding genes, specially designed for rapid comparison of

dozens of closely related prokaryotic genomes. Remarkably, PSP facilitates

functional exploration at the multiple levels by assignments and enrichments of

KO, GO or COG terms. To illustrate this user-friendly tool, we analyzed

Escherichia coli and Bacillus cereus genomes and found that several genes, which

play key roles in human infection and antibiotic resistance, show significant

evidence of positive selection. PSP is freely available to all users without any

login requirement at: http://db-mml.sjtu.edu.cn/PSP/.

CONCLUSIONS: PSP ultimately allows researchers to do genome-scale analysis for

evolutionary selection across multiple prokaryotic genomes rapidly and easily,

and identify the genes undergoing positive selection, which may play key roles in

the interactions of host-pathogen and/or environmental adaptation.

DOI: 10.1186/1471-2164-14-924

PMCID: PMC3882776

PMID: 24373418 [Indexed for MEDLINE]

957. BMC Genomics. 2013 Dec 27;14:922. doi: 10.1186/1471-2164-14-922.

TIPMaP: a web server to establish transcript isoform profiles from reliable

microarray probes.

Chitturi N, Balagannavar G, Chandrashekar DS, Abinaya S, Srini VS, Acharya KK(1).

Author information:

(1)Institute of Bioinformatics and Applied Biotechnology (IBAB), Biotech Park,

Electronic City, Bengaluru (Bangalore) 560 100, Karnataka state, India.

kshitish@ibab.ac.in.

BACKGROUND: Standard 3' Affymetrix gene expression arrays have contributed a

significantly higher volume of existing gene expression data than other

microarray platforms. These arrays were designed to identify differentially

expressed genes, but not their alternatively spliced transcript forms. No

resource can currently identify expression pattern of specific mRNA forms using

these microarray data, even though it is possible to do this.

RESULTS: We report a web server for expression profiling of alternatively spliced

transcripts using microarray data sets from 31 standard 3' Affymetrix arrays for

human, mouse and rat species. The tool has been experimentally validated for

mRNAs transcribed or not-detected in a human disease condition (non-obstructive

azoospermia, a male infertility condition). About 4000 gene expression datasets

were downloaded from a public repository. 'Good probes' with complete coverage

and identity to latest reference transcript sequences were first identified.

Using them, 'Transcript specific probe-clusters' were derived for each platform

and used to identify expression status of possible transcripts. The web server

can lead the user to datasets corresponding to specific tissues, conditions via

identifiers of the microarray studies or hybridizations, keywords, official gene

symbols or reference transcript identifiers. It can identify, in the tissues and

conditions of interest, about 40% of known transcripts as 'transcribed',

'not-detected' or 'differentially regulated'. Corresponding additional

information for probes, genes, transcripts and proteins can be viewed too. We

identified the expression of transcripts in a specific clinical condition and

validated a few of these transcripts by experiments (using reverse transcription

followed by polymerase chain reaction). The experimental observations indicated

higher agreements with the web server results, than contradictions. The tool is

accessible at http://resource.ibab.ac.in/TIPMaP.

CONCLUSION: The newly developed online tool forms a reliable means for

identification of alternatively spliced transcript-isoforms that may be

differentially expressed in various tissues, cell types or physiological

conditions. Thus, by making better use of existing data, TIPMaP avoids the

dependence on precious tissue-samples, in experiments with a goal to establish

expression profiles of alternative splice forms--at least in some cases.

DOI: 10.1186/1471-2164-14-922

PMCID: PMC3884118

PMID: 24373374 [Indexed for MEDLINE]

958. PLoS One. 2013 Dec 23;8(12):e83532. doi: 10.1371/journal.pone.0083532.

eCollection 2013.

NMRDSP: an accurate prediction of protein shape strings from NMR chemical shifts

and sequence data.

Mao W(1), Cong P(1), Wang Z(1), Lu L(1), Zhu Z(1), Li T(1).

Author information:

(1)Department of Chemistry, Tongji University, Shanghai, China.

Shape string is structural sequence and is an extremely important structure

representation of protein backbone conformations. Nuclear magnetic resonance

chemical shifts give a strong correlation with the local protein structure, and

are exploited to predict protein structures in conjunction with computational

approaches. Here we demonstrate a novel approach, NMRDSP, which can accurately

predict the protein shape string based on nuclear magnetic resonance chemical

shifts and structural profiles obtained from sequence data. The NMRDSP uses six

chemical shifts (HA, H, N, CA, CB and C) and eight elements of structure profiles

as features, a non-redundant set (1,003 entries) as the training set, and a

conditional random field as a classification algorithm. For an independent

testing set (203 entries), we achieved an accuracy of 75.8% for S8 (the eight

states accuracy) and 87.8% for S3 (the three states accuracy). This is higher

than only using chemical shifts or sequence data, and confirms that the chemical

shift and the structure profile are significant features for shape string

prediction and their combination prominently improves the accuracy of the

predictor. We have constructed the NMRDSP web server and believe it could be

employed to provide a solid platform to predict other protein structures and

functions. The NMRDSP web server is freely available at

http://cal.tongji.edu.cn/NMRDSP/index.jsp.

DOI: 10.1371/journal.pone.0083532

PMCID: PMC3871590

PMID: 24376713 [Indexed for MEDLINE]

959. PLoS One. 2013 Dec 18;8(12):e82241. doi: 10.1371/journal.pone.0082241.

eCollection 2013.

Online survival analysis software to assess the prognostic value of biomarkers

using transcriptomic data in non-small-cell lung cancer.

Győrffy B(1), Surowiak P(2), Budczies J(3), Lánczky A(1).

Author information:

(1)Research Laboratory of Pediatrics and Nephrology, Hungarian Academy of

Sciences, Budapest, Hungary. (2)Department of Histology and Embryology, Wroclaw

Medical University, Wrocław, Poland. (3)Institut für Pathologie, Charité -

Universitätsmedizin Berlin, Berlin, Germany.

Erratum in

PLoS One. 2014;9(10):e111842.

In the last decade, optimized treatment for non-small cell lung cancer had lead

to improved prognosis, but the overall survival is still very short. To further

understand the molecular basis of the disease we have to identify biomarkers

related to survival. Here we present the development of an online tool suitable

for the real-time meta-analysis of published lung cancer microarray datasets to

identify biomarkers related to survival. We searched the caBIG, GEO and TCGA

repositories to identify samples with published gene expression data and survival

information. Univariate and multivariate Cox regression analysis, Kaplan-Meier

survival plot with hazard ratio and logrank P value are calculated and plotted in

R. The complete analysis tool can be accessed online at: www.kmplot.com/lung. All

together 1,715 samples of ten independent datasets were integrated into the

system. As a demonstration, we used the tool to validate 21 previously published

survival associated biomarkers. Of these, survival was best predicted by CDK1

(p<1E-16), CD24 (p<1E-16) and CADM1 (p = 7E-12) in adenocarcinomas and by CCNE1

(p = 2.3E-09) and VEGF (p = 3.3E-10) in all NSCLC patients. Additional genes

significantly correlated to survival include RAD51, CDKN2A, OPN, EZH2, ANXA3,

ADAM28 and ERCC1. In summary, we established an integrated database and an online

tool capable of uni- and multivariate analysis for in silico validation of new

biomarker candidates in non-small cell lung cancer.

DOI: 10.1371/journal.pone.0082241

PMCID: PMC3867325

PMID: 24367507 [Indexed for MEDLINE]

960. Bioinformatics. 2013 Dec 15;29(24):3204-10. doi: 10.1093/bioinformatics/btt558.

Epub 2013 Sep 27.

STAR: an integrated solution to management and visualization of sequencing data.

Wang T(1), Liu J, Shen L, Tonti-Filippini J, Zhu Y, Jia H, Lister R, Whitaker JW,

Ecker JR, Millar AH, Ren B, Wang W.

Author information:

(1)Department of Chemistry and Biochemistry, University of California, San Diego,

CA 92093, USA, Department of Neuroscience, Icahn School of Medicine at Mount

Sinai, New York, NY 10029, USA, The ARC Centre of Excellence in Plant Energy

Biology, The University of Western Australia, Crawley, Western Australia 6009,

Australia, Key Laboratory for Symbolic Computation and Knowledge Engineering of

Ministry of Education, College of Computer Science and Technology, Jilin

University, Changchun 130012, China, Genomic Analysis Laboratory, The Salk

Institute for Biological Studies, La Jolla, CA 92037, USA, Department of Cellular

and Molecular Medicine, University of California, San Diego, CA 92093, USA and

Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA.

MOTIVATION: Easily visualization of complex data features is a necessary step to

conduct studies on next-generation sequencing (NGS) data. We developed STAR, an

integrated web application that enables online management, visualization and

track-based analysis of NGS data.

RESULTS: STAR is a multilayer web service system. On the client side, STAR

leverages JavaScript, HTML5 Canvas and asynchronous communications to deliver a

smoothly scrolling desktop-like graphical user interface with a suite of

in-browser analysis tools that range from providing simple track configuration

controls to sophisticated feature detection within datasets. On the server side,

STAR supports private session state retention via an account management system

and provides data management modules that enable collection, visualization and

analysis of third-party sequencing data from the public domain with over

thousands of tracks hosted to date. Overall, STAR represents a next-generation

data exploration solution to match the requirements of NGS data, enabling both

intuitive visualization and dynamic analysis of data.

AVAILABILITY AND IMPLEMENTATION: STAR browser system is freely available on the

web at http://wanglab.ucsd.edu/star/browser and

https://github.com/angell1117/STAR-genome-browser.

DOI: 10.1093/bioinformatics/btt558

PMCID: PMC3842760

PMID: 24078702 [Indexed for MEDLINE]

961. Bioinformatics. 2013 Dec 15;29(24):3232-4. doi: 10.1093/bioinformatics/btt562.

Epub 2013 Sep 26.

INVEX--a web-based tool for integrative visualization of expression data.

Xia J(1), Lyle NH, Mayer ML, Pena OM, Hancock RE.

Author information:

(1)Department of Microbiology and Immunology, University of British Columbia,

Vancouver, Canada V6T 1Z3, Centre for Microbial Diseases and Immunity Research,

University of British Columbia, Vancouver, Canada V6T 1Z4 and Wellcome Trust

Sanger Institute, Hinxton CB10 1SA, UK.

SUMMARY: Gene expression or metabolomics data generated from clinical settings

are often associated with multiple metadata (i.e. diagnosis, genotype, gender,

etc.). It is of great interest to analyze and to visualize the data in these

contexts. Here, we introduce INVEX-a novel web-based tool that integrates the

server-side capabilities for data analysis with the browse-based technology for

data visualization. INVEX has two key features: (i) flexible differential

expression analysis for a wide variety of experimental designs; and (ii)

interactive visualization within the context of metadata and biological

annotations. INVEX has built-in support for gene/metabolite annotation and a

fully functional heatmap builder.

AVAILABILITY AND IMPLEMENTATION: Freely available at http://www.invex.ca.

DOI: 10.1093/bioinformatics/btt562

PMCID: PMC3842763

PMID: 24078684 [Indexed for MEDLINE]

962. Bioinformatics. 2013 Dec 15;29(24):3230-1. doi: 10.1093/bioinformatics/btt561.

Epub 2013 Sep 26.

DockAFM: benchmarking protein structures by docking under AFM topographs.

Chaves RC(1), Pellequer JL.

Author information:

(1)CEA, iBEB, Service de Biochimie et Toxicologie Nucléaire, F-30207 Bagnols sur

Cèze, France.

Proteins can adopt a variety of conformations. We present a simple server for

scoring the agreement between 3D atomic structures and experimental envelopes

obtained by atomic force microscopy. Three different structures of

immunoglobulins (IgG) or blood coagulation factor V activated were tested and

their agreement with several topographical surfaces was computed. This approach

can be used to test structural variability within a family of

proteins.AVAILABILITY AND IMPLEMENTATION: DockAFM is available at

http://biodev.cea.fr/dockafm.

DOI: 10.1093/bioinformatics/btt561

PMID: 24078683 [Indexed for MEDLINE]

963. Bioinformatics. 2013 Dec 15;29(24):3238-40. doi: 10.1093/bioinformatics/btt559.

Epub 2013 Sep 25.

WebGLORE: a web service for Grid LOgistic REgression.

Jiang W(1), Li P, Wang S, Wu Y, Xue M, Ohno-Machado L, Jiang X.

Author information:

(1)Division of Biomedical Informatics, University of California, San Diego, La

Jolla, CA 92093, USA and Department of Electronic Engineering, Shanghai Jiao Tong

University, Shanghai 200240, China.

WebGLORE is a free web service that enables privacy-preserving construction of a

global logistic regression model from distributed datasets that are sensitive. It

only transfers aggregated local statistics (from participants) through Hypertext

Transfer Protocol Secure to a trusted server, where the global model is

synthesized. WebGLORE seamlessly integrates AJAX, JAVA Applet/Servlet and PHP

technologies to provide an easy-to-use web service for biomedical researchers to

break down policy barriers during information exchange.AVAILABILITY AND

IMPLEMENTATION: http://dbmi-engine.ucsd.edu/webglore3/. WebGLORE can be used

under the terms of GNU general public license as published by the Free Software

Foundation.

DOI: 10.1093/bioinformatics/btt559

PMCID: PMC3842761

PMID: 24072732 [Indexed for MEDLINE]

964. Bioinformatics. 2013 Dec 15;29(24):3135-42. doi: 10.1093/bioinformatics/btt554.

Epub 2013 Sep 23.

Accurate prediction of bacterial type IV secreted effectors using amino acid

composition and PSSM profiles.

Zou L(1), Nan C, Hu F.

Author information:

(1)Department of Microbiology, College of Basic Medical Sciences, Third Military

Medical University (TMMU), Chongqing 40038, China and Department of Tuberculosis,

Institute of Infectious TB Prevention, Third Hospital of PLA, Baoji, Shanxi

721006, China.

MOTIVATION: Various human pathogens secret effector proteins into hosts cells via

the type IV secretion system (T4SS). These proteins play important roles in the

interaction between bacteria and hosts. Computational methods for T4SS effector

prediction have been developed for screening experimental targets in several

isolated bacterial species; however, widely applicable prediction approaches are

still unavailable

RESULTS: In this work, four types of distinctive features, namely, amino acid

composition, dipeptide composition, .position-specific scoring matrix composition

and auto covariance transformation of position-specific scoring matrix, were

calculated from primary sequences. A classifier, T4EffPred, was developed using

the support vector machine with these features and their different combinations

for effector prediction. Various theoretical tests were performed in a newly

established dataset, and the results were measured with four indexes. We

demonstrated that T4EffPred can discriminate IVA and IVB effectors in benchmark

datasets with positive rates of 76.7% and 89.7%, respectively. The overall

accuracy of 95.9% shows that the present method is accurate for distinguishing

the T4SS effector in unidentified sequences. A classifier ensemble was designed

to synthesize all single classifiers. Notable performance improvement was

observed using this ensemble system in benchmark tests. To demonstrate the

model's application, a genome-scale prediction of effectors was performed in

Bartonella henselae, an important zoonotic pathogen. A number of putative

candidates were distinguished.

AVAILABILITY: A web server implementing the prediction method and the source code

are both available at http://bioinfo.tmmu.edu.cn/T4EffPred.

DOI: 10.1093/bioinformatics/btt554

PMID: 24064423 [Indexed for MEDLINE]

965. Bioinformatics. 2013 Dec 15;29(24):3225-6. doi: 10.1093/bioinformatics/btt545.

Epub 2013 Sep 18.

miREval 2.0: a web tool for simple microRNA prediction in genome sequences.

Gao D(1), Middleton R, Rasko JE, Ritchie W.

Author information:

(1)Bioinformatics Laboratory, Centenary Institute, Gene and Stem Cell Therapy

Program, Centenary Institute, University of Sydney, Sydney, New South Wales,

Australia and Cell and Molecular Therapies, Royal Prince Alfred Hospital,

Camperdown, New South Wales 2050, Australia.

RESULT: We have developed miREval 2.0, an online tool that can simultaneously

search up to 100 sequences for novel microRNAs (miRNAs) in multiple organisms.

miREval 2.0 uses multiple published in silico approaches to detect miRNAs in

sequences of interest. This tool can be used to discover miRNAs from DNA

sequences or to validate candidates from sequencing data.

AVAILABILITY: http://mimirna.centenary.org.au/mireval/.

DOI: 10.1093/bioinformatics/btt545

PMID: 24048357 [Indexed for MEDLINE]

966. BMC Struct Biol. 2013 Dec 13;13:32. doi: 10.1186/1472-6807-13-32.

BioSuper: a web tool for the superimposition of biomolecules and assemblies with

rotational symmetry.

Rueda M, Orozco M, Totrov M, Abagyan R(1).

Author information:

(1)Skaggs School of Pharmacy and Pharmaceutical Sciences, University of

California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.

rabagyan@ucsd.edu.

BACKGROUND: Most of the proteins in the Protein Data Bank (PDB) are oligomeric

complexes consisting of two or more subunits that associate by rotational or

helical symmetries. Despite the myriad of superimposition tools in the

literature, we could not find any able to account for rotational symmetry and

display the graphical results in the web browser.

RESULTS: BioSuper is a free web server that superimposes and calculates the root

mean square deviation (RMSD) of protein complexes displaying rotational symmetry.

To the best of our knowledge, BioSuper is the first tool of its kind that

provides immediate interactive visualization of the graphical results in the

browser, biomolecule generator capabilities, different levels of atom selection,

sequence-dependent and structure-based superimposition types, and is the only web

tool that takes into account the equivalence of atoms in side chains displaying

symmetry ambiguity. BioSuper uses ICM program functionality as a core for the

superimpositions and displays the results as text, HTML tables and 3D interactive

molecular objects that can be visualized in the browser or in Android and iOS

platforms with a free plugin.

CONCLUSIONS: BioSuper is a fast and functional tool that allows for pairwise

superimposition of proteins and assemblies displaying rotational symmetry. The

web server was created after our own frustration when attempting to superimpose

flexible oligomers. We strongly believe that its user-friendly and functional

design will be of great interest for structural and computational biologists who

need to superimpose oligomeric proteins (or any protein). BioSuper web server is

freely available to all users at http://ablab.ucsd.edu/BioSuper.

DOI: 10.1186/1472-6807-13-32

PMCID: PMC3924234

PMID: 24330655 [Indexed for MEDLINE]

967. Genome Biol. 2013 Dec 13;14(12):R134. doi: 10.1186/gb-2013-14-12-r134.

TRAPID: an efficient online tool for the functional and comparative analysis of

de novo RNA-Seq transcriptomes.

Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K.

Transcriptome analysis through next-generation sequencing technologies allows the

generation of detailed gene catalogs for non-model species, at the cost of new

challenges with regards to computational requirements and bioinformatics

expertise. Here, we present TRAPID, an online tool for the fast and efficient

processing of assembled RNA-Seq transcriptome data, developed to mitigate these

challenges. TRAPID offers high-throughput open reading frame detection,

frameshift correction and includes a functional, comparative and phylogenetic

toolbox, making use of 175 reference proteomes. Benchmarking and comparison

against state-of-the-art transcript analysis tools reveals the efficiency and

unique features of the TRAPID system. TRAPID is freely available at

http://bioinformatics.psb.ugent.be/webtools/trapid/.

DOI: 10.1186/gb-2013-14-12-r134

PMCID: PMC4053847

PMID: 24330842 [Indexed for MEDLINE]

968. J Transl Med. 2013 Dec 11;11:305. doi: 10.1186/1479-5876-11-305.

VIRsiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting

human viruses.

Qureshi A, Thakur N, Kumar M(1).

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Council of

Scientific and Industrial Research, Sector 39-A, Chandigarh 160036, India.

manojk@imtech.res.in.

BACKGROUND: Selection of effective viral siRNA is an indispensable step in the

development of siRNA based antiviral therapeutics. Despite immense potential, a

viral siRNA efficacy prediction algorithm is still not available. Moreover,

performances of the existing general mammalian siRNA efficacy predictors are not

satisfactory for viral siRNAs. Therefore, we have developed "VIRsiRNApred" a

support vector machine (SVM) based method for predicting the efficacy of viral

siRNA.

METHODS: In the present study, we have employed a new dataset of 1725 viral

siRNAs with experimentally verified quantitative efficacies tested under

heterogeneous experimental conditions and targeting as many as 37 important human

viruses including HIV, Influenza, HCV, HBV, SARS etc. These siRNAs were divided

into training (T1380) and validation (V345) datasets. Important siRNA sequence

features including mono to penta nucleotide frequencies, binary pattern,

thermodynamic properties and secondary structure were employed for model

development.

RESULTS: During 10-fold cross validation on T1380 using hybrid approach, we

achieved a maximum Pearson Correlation Coefficient (PCC) of 0.55 between

predicted and actual efficacy of viral siRNAs. On V345 independent dataset, our

best model achieved a maximum correlation of 0.50 while existing general siRNA

prediction methods showed PCC from 0.05 to 0.18. However, using leave one out

cross validation PCC was improved to 0.58 and 0.55 on training and validation

datasets respectively. SVM performed better than other machine learning

techniques used like ANN, KNN and REP Tree.

CONCLUSION: VIRsiRNApred is the first algorithm for predicting inhibition

efficacy of viral siRNAs which is developed using experimentally verified viral

siRNAs. We hope this algorithm would be useful in predicting highly potent viral

siRNA to aid siRNA based antiviral therapeutics development. The web server is

freely available at http://crdd.osdd.net/servers/virsirnapred/.

DOI: 10.1186/1479-5876-11-305

PMCID: PMC3878835

PMID: 24330765 [Indexed for MEDLINE]

969. BMC Genet. 2013 Dec 9;14:118. doi: 10.1186/1471-2156-14-118.

Development of a model webserver for breed identification using microsatellite

DNA marker.

Iquebal MA, Sarika, Dhanda SK, Arora V, Dixit SP, Raghava GP, Rai A, Kumar D(1).

Author information:

(1)Centre for Agricultural Bioinformatics, Indian Agricultural Statistics

Research Institute, Library Avenue, PUSA, New Delhi 110012, India.

dineshkumarbhu@gmail.com.

BACKGROUND: Identification of true to breed type animal for conservation purpose

is imperative. Breed dilution is one of the major problems in sustainability

except cases of commercial crossbreeding under controlled condition. Breed

descriptor has been developed to identify breed but such descriptors cover only

"pure breed" or true to the breed type animals excluding undefined or admixture

population. Moreover, in case of semen, ova, embryo and breed product, the breed

cannot be identified due to lack of visible phenotypic descriptors. Advent of

molecular markers like microsatellite and SNP have revolutionized breed

identification from even small biological tissue or germplasm. Microsatellite DNA

marker based breed assignments has been reported in various domestic animals.

Such methods have limitations viz. non availability of allele data in public

domain, thus each time all reference breed has to be genotyped which is neither

logical nor economical. Even if such data is available but computational methods

needs expertise of data analysis and interpretation.

RESULTS: We found Bayesian Networks as best classifier with highest accuracy of

98.7% using 51850 reference allele data generated by 25 microsatellite loci on 22

goat breed population of India. The FST values in the study were seen to be low

ranging from 0.051 to 0.297 and overall genetic differentiation of 13.8%,

suggesting more number of loci needed for higher accuracy. We report here world's

first model webserver for breed identification using microsatellite DNA markers

freely accessible at http://cabin.iasri.res.in/gomi/.

CONCLUSION: Higher number of loci is required due to less differentiable

population and large number of breeds taken in this study. This server will

reduce the cost with computational ease. This methodology can be a model for

various other domestic animal species as a valuable tool for conservation and

breed improvement programmes.

DOI: 10.1186/1471-2156-14-118

PMCID: PMC3890620

PMID: 24320218 [Indexed for MEDLINE]

970. PLoS One. 2013 Dec 6;8(12):e82210. doi: 10.1371/journal.pone.0082210. eCollection

2013.

INDIGO - INtegrated data warehouse of microbial genomes with examples from the

red sea extremophiles.

Alam I(1), Antunes A, Kamau AA, Ba Alawi W, Kalkatawi M, Stingl U, Bajic VB.

Author information:

(1)Computational Bioscience Research Center (CBRC), King Abdullah University of

Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia.

BACKGROUND: The next generation sequencing technologies substantially increased

the throughput of microbial genome sequencing. To functionally annotate newly

sequenced microbial genomes, a variety of experimental and computational methods

are used. Integration of information from different sources is a powerful

approach to enhance such annotation. Functional analysis of microbial genomes,

necessary for downstream experiments, crucially depends on this annotation but it

is hampered by the current lack of suitable information integration and

exploration systems for microbial genomes.

RESULTS: We developed a data warehouse system (INDIGO) that enables the

integration of annotations for exploration and analysis of newly sequenced

microbial genomes. INDIGO offers an opportunity to construct complex queries and

combine annotations from multiple sources starting from genomic sequence to

protein domain, gene ontology and pathway levels. This data warehouse is aimed at

being populated with information from genomes of pure cultures and uncultured

single cells of Red Sea bacteria and Archaea. Currently, INDIGO contains

information from Salinisphaera shabanensis, Haloplasma contractile, and

Halorhabdus tiamatea - extremophiles isolated from deep-sea anoxic brine lakes of

the Red Sea. We provide examples of utilizing the system to gain new insights

into specific aspects on the unique lifestyle and adaptations of these organisms

to extreme environments.

CONCLUSIONS: We developed a data warehouse system, INDIGO, which enables

comprehensive integration of information from various resources to be used for

annotation, exploration and analysis of microbial genomes. It will be regularly

updated and extended with new genomes. It is aimed to serve as a resource

dedicated to the Red Sea microbes. In addition, through INDIGO, we provide our

Automatic Annotation of Microbial Genomes (AAMG) pipeline. The INDIGO web server

is freely available at http://www.cbrc.kaust.edu.sa/indigo.

DOI: 10.1371/journal.pone.0082210

PMCID: PMC3855842

PMID: 24324765 [Indexed for MEDLINE]

971. PLoS One. 2013 Dec 4;8(12):e80660. doi: 10.1371/journal.pone.0080660. eCollection

2013.

PMS: a panoptic motif search tool.

Dinh H(1), Rajasekaran S.

Author information:

(1)Computer Science and Engineering Department, University of Connecticut,

Storrs, Connecticut, United States of America.

BACKGROUND: Identification of DNA/Protein motifs is a crucial problem for

biologists. Computational techniques could be of great help in this

identification. In this direction, many computational models for motifs have been

proposed in the literature.

METHODS: One such important model is the (l,d) motif model. In this paper we

describe a motif search web tool that predominantly employs this motif model.

This web tool exploits the state-of-the art algorithms for solving the (l,d)

motif search problem.

RESULTS: The online tool has been helping scientists identify many unknown

motifs. Many of our predictions have been successfully verified as well. We hope

that this paper will expose this crucial tool to many more scientists.

AVAILABILITY AND REQUIREMENTS: Project name: PMS--Panoptic Motif Search Tool.

Project home page: http://pms.engr.uconn.edu or http://motifsearch.com. Licence:

PMS tools will be readily available to any scientist wishing to use it for

non-commercial purposes, without restrictions. The online tool is freely

available without login.

DOI: 10.1371/journal.pone.0080660

PMCID: PMC3851466

PMID: 24324619 [Indexed for MEDLINE]

972. PLoS One. 2013 Dec 2;8(12):e81979. doi: 10.1371/journal.pone.0081979. eCollection

2013.

Structural and functional analysis of human SOD1 in amyotrophic lateral

sclerosis.

Moreira LG(1), Pereira LC, Drummond PR, De Mesquita JF.

Author information:

(1)Bioinformatics and Computational Biology Group, Department of Genetics and

Molecular Biology, Federal University of Rio de Janeiro State (UNIRIO), Rio de

Janeiro, Brazil.

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease with

familial inheritance (fALS) in 5% to 10% of cases; 25% of those are caused by

mutations in the superoxide dismutase 1 (SOD1) protein. More than 100 mutations

in the SOD1 gene have been associated with fALS, altering the geometry of the

active site, protein folding and the interaction between monomers. We performed a

functional analysis of non-synonymous single nucleotide polymorphisms (nsSNPs) in

124 fALS SOD1 mutants. Eleven different algorithms were used to estimate the

functional impact of the replacement of one amino acid on protein structure:

SNPs&GO, PolyPhen-2, SNAP, PMUT, Sift, PhD-SNP, nsSNPAnalyzer, TANGO, WALTZ,

LIMBO and FoldX. For the structural analysis, theoretical models of 124 SNPs of

SOD1 were created by comparative modeling using the MHOLline workflow, which

includes Modeller and Procheck. Models were aligned with the native protein by

the TM-align algorithm. A human-curated database was developed using the server

side include in Java, JMOL. The results of this functional analysis indicate that

the majority of the 124 natural mutants are harmful to the protein structure and

thus corroborate the correlation between the reported mutations and fALS. In the

structural analysis, all models showed conformational changes when compared to

wild-type SOD1, and the degree of structural alignment varied between them. The

SOD1 database converge structural and functional analyses of SOD1; it is a vast

resource for the molecular analysis of amyotrophic lateral sclerosis, which

allows the user to expand his knowledge on the molecular basis of the disease.

The SOD1 database is available at http://bioinfogroup.com/database.

DOI: 10.1371/journal.pone.0081979

PMCID: PMC3846731

PMID: 24312616 [Indexed for MEDLINE]

973. Acta Crystallogr D Biol Crystallogr. 2013 Dec;69(Pt 12):2395-402. doi:

10.1107/S0907444913022294. Epub 2013 Nov 19.

LigSearch: a knowledge-based web server to identify likely ligands for a protein

target.

de Beer TA(1), Laskowski RA, Duban ME, Chan AW, Anderson WF, Thornton JM.

Author information:

(1)European Molecular Biology Laboratory, European Bioinformatics Institute

(EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England.

Identifying which ligands might bind to a protein before crystallization trials

could provide a significant saving in time and resources. LigSearch, a web server

aimed at predicting ligands that might bind to and stabilize a given protein, has

been developed. Using a protein sequence and/or structure, the system searches

against a variety of databases, combining available knowledge, and provides a

clustered and ranked output of possible ligands. LigSearch can be accessed at

http://www.ebi.ac.uk/thornton-srv/databases/LigSearch.

DOI: 10.1107/S0907444913022294

PMCID: PMC3852652

PMID: 24311580 [Indexed for MEDLINE]

974. BMC Bioinformatics. 2013 Dec 1;14:346. doi: 10.1186/1471-2105-14-346.

Disulfide by Design 2.0: a web-based tool for disulfide engineering in proteins.

Craig DB, Dombkowski AA(1).

Author information:

(1)Department of Pediatrics, Wayne State University School of Medicine, Detroit,

Michigan 48201, USA. domski@wayne.edu.

BACKGROUND: Disulfide engineering is an important biotechnological tool that has

advanced a wide range of research. The introduction of novel disulfide bonds into

proteins has been used extensively to improve protein stability, modify

functional characteristics, and to assist in the study of protein dynamics.

Successful use of this technology is greatly enhanced by software that can

predict pairs of residues that will likely form a disulfide bond if mutated to

cysteines.

RESULTS: We had previously developed and distributed software for this purpose:

Disulfide by Design (DbD). The original DbD program has been widely used;

however, it has a number of limitations including a Windows platform dependency.

Here, we introduce Disulfide by Design 2.0 (DbD2), a web-based,

platform-independent application that significantly extends functionality,

visualization, and analysis capabilities beyond the original program. Among the

enhancements to the software is the ability to analyze the B-factor of protein

regions involved in predicted disulfide bonds. Importantly, this feature

facilitates the identification of potential disulfides that are not only likely

to form but are also expected to provide improved thermal stability to the

protein.

CONCLUSIONS: DbD2 provides platform-independent access and significantly extends

the original functionality of DbD. A web server hosting DbD2 is provided at

http://cptweb.cpt.wayne.edu/DbD2/.

DOI: 10.1186/1471-2105-14-346

PMCID: PMC3898251

PMID: 24289175 [Indexed for MEDLINE]

975. Hum Mutat. 2013 Dec;34(12):1606-10. doi: 10.1002/humu.22444. Epub 2013 Oct 10.

RRBS-analyser: a comprehensive web server for reduced representation bisulfite

sequencing data analysis.

Wang T(1), Liu Q, Li X, Wang X, Li J, Zhu X, Sun ZS, Wu J.

Author information:

(1)Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, 325035,

China.

In reduced representation bisulfite sequencing (RRBS), genomic DNA is digested

with the restriction enzyme and then subjected to next-generation sequencing,

which enables detection and quantification of DNA methylation at whole-genome

scale with low cost. However, the data processing, interpretation, and analysis

of the huge amounts of data generated pose a bioinformatics challenge. We

developed RRBS-Analyser, a comprehensive genome-scale DNA methylation analysis

server based on RRBS data. RRBS-Analyser can assess sequencing quality, generate

detailed statistical information, align the bisulfite-treated short reads to

reference genome, identify and annotate the methylcytosines (5mCs) and associate

them with different genomic features in CG, CHG, and CHH content. RRBS-Analyser

supports detection, annotation, and visualization of differentially methylated

regions (DMRs) for multiple samples from nine reference organisms. Moreover,

RRBS-Analyser provides researchers with detailed annotation of DMR-containing

genes, which will greatly aid subsequent studies. The input of RRBS-Analyser can

be raw FASTQ reads, generic SAM format, or self-defined format containing

individual 5mC sites. RRBS-Analyser can be widely used by researchers wanting to

unravel the complexities of DNA methylome in the epigenetic community.

RRBS-Analyser is freely available at http://122.228.158.106/RRBSAnalyser/.

© 2013 WILEY PERIODICALS, INC.

DOI: 10.1002/humu.22444

PMID: 24106010 [Indexed for MEDLINE]

976. J Bioinform Comput Biol. 2013 Dec;11(6):1343011. doi: 10.1142/S0219720013430117.

Epub 2013 Dec 11.

Automatic phylogenetic classification of bacterial beta-lactamase sequences

including structural and antibiotic substrate preference information.

Ma J(1), Eisenhaber F, Maurer-Stroh S.

Author information:

(1)Bioinformatics Institute (BII), Agency for Science, Technology and Research

(A\*STAR), 30 Biopolis Street, #07-01 Matrix, Singapore 138671, Singapore.

Beta lactams comprise the largest and still most effective group of antibiotics,

but bacteria can gain resistance through different beta lactamases that can

degrade these antibiotics. We developed a user friendly tree building web server

that allows users to assign beta lactamase sequences to their respective

molecular classes and subclasses. Further clinically relevant information

includes if the gene is typically chromosomal or transferable through plasmids as

well as listing the antibiotics which the most closely related reference

sequences are known to target and cause resistance against. This web server can

automatically build three phylogenetic trees: the first tree with closely related

sequences from a Tachyon search against the NCBI nr database, the second tree

with curated reference beta lactamase sequences, and the third tree built

specifically from substrate binding pocket residues of the curated reference beta

lactamase sequences. We show that the latter is better suited to recover

antibiotic substrate assignments through nearest neighbor annotation transfer.

The users can also choose to build a structural model for the query sequence and

view the binding pocket residues of their query relative to other beta lactamases

in the sequence alignment as well as in the 3D structure relative to bound

antibiotics. This web server is freely available at

http://blac.bii.a-star.edu.sg/.

DOI: 10.1142/S0219720013430117

PMID: 24372040 [Indexed for MEDLINE]

977. J Bioinform Comput Biol. 2013 Dec;11(6):1343009. doi: 10.1142/S0219720013430099.

Epub 2013 Dec 11.

Ab initio human miRNA and pre-miRNA prediction.

Titov II(1), Vorozheykin PS.

Author information:

(1)Institute of Cytology and Genetics, SB RAS, 10 Lavrentyev Avenue, Novosibirsk

630090, Russian Federation , Novosibirsk State University, 2 Pirogov Street,

Novosibirsk 630090, Russian Federation.

MicroRNAs (miRNAs) are small single-stranded noncoding RNAs that play an

important role in post-transcriptional regulation of gene expression. In this

paper, we present a web server for ab initio prediction of the human miRNAs and

their precursors. The prediction methods are based on the hidden Markov Models

and the context-structural characteristics. By taking into account the identified

patterns of primary and secondary structures of the pre-miRNAs, a new HMM model

is proposed and the existing context-structural Markov model is modified. The

evaluation of the method performance has shown that it can accurately predict

novel human miRNAs. Comparing with the existing methods we demonstrate that our

method has a higher prediction quality both for human pre-miRNAs and miRNAs. The

models have also showed good results in the prediction of the mouse miRNAs. The

web server is available at http://wwwmgs.bionet.nsc.ru/mgs/programs/rnaanalys

(mirror http://miRNA.at.nsu.ru ).

DOI: 10.1142/S0219720013430099

PMID: 24372038 [Indexed for MEDLINE]

978. RNA. 2013 Dec;19(12):1605-16. doi: 10.1261/rna.039834.113. Epub 2013 Oct 21.

LigandRNA: computational predictor of RNA-ligand interactions.

Philips A, Milanowska K, Lach G, Bujnicki JM.

RNA molecules have recently become attractive as potential drug targets due to

the increased awareness of their importance in key biological processes. The

increase of the number of experimentally determined RNA 3D structures enabled

structure-based searches for small molecules that can specifically bind to

defined sites in RNA molecules, thereby blocking or otherwise modulating their

function. However, as of yet, computational methods for structure-based docking

of small molecule ligands to RNA molecules are not as well established as

analogous methods for protein-ligand docking. This motivated us to create

LigandRNA, a scoring function for the prediction of RNA-small molecule

interactions. Our method employs a grid-based algorithm and a knowledge-based

potential derived from ligand-binding sites in the experimentally solved

RNA-ligand complexes. As an input, LigandRNA takes an RNA receptor file and a

file with ligand poses. As an output, it returns a ranking of the poses according

to their score. The predictive power of LigandRNA favorably compares to five

other publicly available methods. We found that the combination of LigandRNA and

Dock6 into a "meta-predictor" leads to further improvement in the identification

of near-native ligand poses. The LigandRNA program is available free of charge as

a web server at http://ligandrna.genesilico.pl.

DOI: 10.1261/rna.039834.113

PMCID: PMC3860260

PMID: 24145824 [Indexed for MEDLINE]

979. BMC Bioinformatics. 2013 Nov 29;14:345. doi: 10.1186/1471-2105-14-345.

wKinMut: an integrated tool for the analysis and interpretation of mutations in

human protein kinases.

Izarzugaza JM(1), Vazquez M, del Pozo A, Valencia A.

Author information:

(1)Structural Biology and BioComputing Programme, Spanish National Cancer

Research Centre (CNIO), C/Melchor Fernandez Almagro, 3, E-28029 Madrid, Spain.

josemgizarzugaza@cbs.dtu.dk.

BACKGROUND: Protein kinases are involved in relevant physiological functions and

a broad number of mutations in this superfamily have been reported in the

literature to affect protein function and stability. Unfortunately, the

exploration of the consequences on the phenotypes of each individual mutation

remains a considerable challenge.

RESULTS: The wKinMut web-server offers direct prediction of the potential

pathogenicity of the mutations from a number of methods, including our recently

developed prediction method based on the combination of information from a range

of diverse sources, including physicochemical properties and functional

annotations from FireDB and Swissprot and kinase-specific characteristics such as

the membership to specific kinase groups, the annotation with disease-associated

GO terms or the occurrence of the mutation in PFAM domains, and the relevance of

the residues in determining kinase subfamily specificity from S3Det. This

predictor yields interesting results that compare favourably with other methods

in the field when applied to protein kinases.Together with the predictions,

wKinMut offers a number of integrated services for the analysis of mutations.

These include: the classification of the kinase, information about associations

of the kinase with other proteins extracted from iHop, the mapping of the

mutations onto PDB structures, pathogenicity records from a number of databases

and the classification of mutations in large-scale cancer studies. Importantly,

wKinMut is connected with the SNP2L system that extracts mentions of mutations

directly from the literature, and therefore increases the possibilities of

finding interesting functional information associated to the studied mutations.

CONCLUSIONS: wKinMut facilitates the exploration of the information available

about individual mutations by integrating prediction approaches with the

automatic extraction of information from the literature (text mining) and several

state-of-the-art databases.wKinMut has been used during the last year for the

analysis of the consequences of mutations in the context of a number of cancer

genome projects, including the recent analysis of Chronic Lymphocytic Leukemia

cases and is publicly available at http://wkinmut.bioinfo.cnio.es.

DOI: 10.1186/1471-2105-14-345

PMCID: PMC3879071

PMID: 24289158 [Indexed for MEDLINE]

980. BMC Bioinformatics. 2013 Nov 28;14:342. doi: 10.1186/1471-2105-14-342.

A novel approach for protein subcellular location prediction using amino acid

exposure.

Mer AS(1), Andrade-Navarro MA.

Author information:

(1)Computational Biology and Data Mining, Max Delbrück Center for Molecular

Medicine, Robert-Rössle-Str, 10, Berlin 13125, Germany.

miguel.andrade@mdc-berlin.de.

BACKGROUND: Proteins perform their functions in associated cellular locations.

Therefore, the study of protein function can be facilitated by predictions of

protein location. Protein location can be predicted either from the sequence of a

protein alone by identification of targeting peptide sequences and motifs, or by

homology to proteins of known location. A third approach, which is complementary,

exploits the differences in amino acid composition of proteins associated to

different cellular locations, and can be useful if motif and homology information

are missing. Here we expand this approach taking into account amino acid

composition at different levels of amino acid exposure.

RESULTS: Our method has two stages. For stage one, we trained multiple Support

Vector Machines (SVMs) to score eukaryotic protein sequences for membership to

each of three categories: nuclear, cytoplasmic and extracellular, plus extra

category nucleocytoplasmic, accounting for the fact that a large number of

proteins shuttles between those two locations. In stage two we use an artificial

neural network (ANN) to propose a category from the scores given to the four

locations in stage one. The method reaches an accuracy of 68% when using as input

3D-derived values of amino acid exposure. Calibration of the method using

predicted values of amino acid exposure allows classifying proteins without

3D-information with an accuracy of 62% and discerning proteins in different

locations even if they shared high levels of identity.

CONCLUSIONS: In this study we explored the relationship between residue exposure

and protein subcellular location. We developed a new algorithm for subcellular

location prediction that uses residue exposure signatures. Our algorithm uses a

novel approach to address the multiclass classification problem. The algorithm is

implemented as web server 'NYCE' and can be accessed at

http://cbdm.mdc-berlin.de/~amer/nyce.

DOI: 10.1186/1471-2105-14-342

PMCID: PMC4219330

PMID: 24283794 [Indexed for MEDLINE]

981. PLoS One. 2013 Nov 26;8(11):e80493. doi: 10.1371/journal.pone.0080493.

eCollection 2013.

Detecting protein candidate fragments using a structural alphabet profile

comparison approach.

Shen Y(1), Picord G, Guyon F, Tuffery P.

Author information:

(1)INSERM, U973, MTi, Paris, France ; Univ Paris Diderot, Sorbonne Paris Cité,

Paris, France.

Predicting accurate fragments from sequence has recently become a critical step

for protein structure modeling, as protein fragment assembly techniques are

presently among the most efficient approaches for de novo prediction. A key step

in these approaches is, given the sequence of a protein to model, the

identification of relevant fragments - candidate fragments - from a collection of

the available 3D structures. These fragments can then be assembled to produce a

model of the complete structure of the protein of interest. The search for

candidate fragments is classically achieved by considering local sequence

similarity using profile comparison, or threading approaches. In the present

study, we introduce a new profile comparison approach that, instead of using

amino acid profiles, is based on the use of predicted structural alphabet

profiles, where structural alphabet profiles contain information related to the

3D local shapes associated with the sequences. We show that structural alphabet

profile-profile comparison can be used efficiently to retrieve accurate

structural fragments, and we introduce a fully new protocol for the detection of

candidate fragments. It identifies fragments specific of each position of the

sequence and of size varying between 6 and 27 amino-acids. We find it outperforms

present state of the art approaches in terms (i) of the accuracy of the fragments

identified, (ii) the rate of true positives identified, while having a high

coverage score. We illustrate the relevance of the approach on complete target

sets of the two previous Critical Assessment of Techniques for Protein Structure

Prediction (CASP) rounds 9 and 10. A web server for the approach is freely

available at http://bioserv.rpbs.univ-paris-diderot.fr/SAFrag.

DOI: 10.1371/journal.pone.0080493

PMCID: PMC3841190

PMID: 24303019 [Indexed for MEDLINE]

982. J Chem Inf Model. 2013 Nov 25;53(11):3097-112. doi: 10.1021/ci400510e. Epub 2013

Nov 11.

Nonlinear scoring functions for similarity-based ligand docking and binding

affinity prediction.

Brylinski M(1).

Author information:

(1)Department of Biological Sciences, Louisiana State University , Baton Rouge,

Louisiana 70803, United States.

A common strategy for virtual screening considers a systematic docking of a large

library of organic compounds into the target sites in protein receptors with

promising leads selected based on favorable intermolecular interactions. Despite

a continuous progress in the modeling of protein-ligand interactions for

pharmaceutical design, important challenges still remain, thus the development of

novel techniques is required. In this communication, we describe eSimDock, a new

approach to ligand docking and binding affinity prediction. eSimDock employs

nonlinear machine learning-based scoring functions to improve the accuracy of

ligand ranking and similarity-based binding pose prediction, and to increase the

tolerance to structural imperfections in the target structures. In large-scale

benchmarking using the Astex/CCDC data set, we show that 53.9% (67.9%) of the

predicted ligand poses have RMSD of <2 Å (<3 Å). Moreover, using binding sites

predicted by recently developed eFindSite, eSimDock models ligand binding poses

with an RMSD of 4 Å for 50.0-39.7% of the complexes at the protein homology level

limited to 80-40%. Simulations against non-native receptor structures, whose mean

backbone rearrangements vary from 0.5 to 5.0 Å Cα-RMSD, show that the ratio of

docking accuracy and the estimated upper bound is at a constant level of ∼0.65.

Pearson correlation coefficient between experimental and predicted by eSimDock Ki

values for a large data set of the crystal structures of protein-ligand complexes

from BindingDB is 0.58, which decreases only to 0.46 when target structures

distorted to 3.0 Å Cα-RMSD are used. Finally, two case studies demonstrate that

eSimDock can be customized to specific applications as well. These encouraging

results show that the performance of eSimDock is largely unaffected by the

deformations of ligand binding regions, thus it represents a practical strategy

for across-proteome virtual screening using protein models. eSimDock is freely

available to the academic community as a Web server at

http://www.brylinski.org/esimdock .

DOI: 10.1021/ci400510e

PMID: 24171431 [Indexed for MEDLINE]

983. J Chem Inf Model. 2013 Nov 25;53(11):2812-9. doi: 10.1021/ci400326p. Epub 2013

Nov 11.

ReactionMap: an efficient atom-mapping algorithm for chemical reactions.

Fooshee D(1), Andronico A, Baldi P.

Author information:

(1)Institute for Genomics and Bioinformatics and School of Information and

Computer Sciences, University of California , Irvine, California 92697, United

States.

Large databases of chemical reactions provide new data-mining opportunities and

challenges. Key challenges result from the imperfect quality of the data and the

fact that many of these reactions are not properly balanced or atom-mapped. Here,

we describe ReactionMap, an efficient atom-mapping algorithm. Our approach uses a

combination of maximum common chemical subgraph search and minimization of an

assignment cost function derived empirically from training data. We use a set of

over 259,000 balanced atom-mapped reactions from the SPRESI commercial database

to train the system, and we validate it on random sets of 1000 and 17,996

reactions sampled from this pool. These large test sets represent a broad range

of chemical reaction types, and ReactionMap correctly maps about 99% of the atoms

and about 96% of the reactions, with a mean time per mapping of 2 s. Most

correctly mapped reactions are mapped with high confidence. Mapping accuracy

compares favorably with ChemAxon's AutoMapper, versions 5 and 6.1, and the DREAM

Web tool. These approaches correctly map 60.7%, 86.5%, and 90.3% of the

reactions, respectively, on the same data set. A ReactionMap server is available

on the ChemDB Web portal at http://cdb.ics.uci.edu .

DOI: 10.1021/ci400326p

PMID: 24160861 [Indexed for MEDLINE]

984. J Chem Inf Model. 2013 Nov 25;53(11):2820-8. doi: 10.1021/ci400432a. Epub 2013

Nov 1.

Exploring the biologically relevant chemical space for drug discovery.

Deng ZL(1), Du CX, Li X, Hu B, Kuang ZK, Wang R, Feng SY, Zhang HY, Kong DX.

Author information:

(1)State Key Laboratory of Agricultural Microbiology, ‡Center for Bioinformatics,

College of Life Science and Technology, and §National Key Laboratory of Crop

Genetic Improvement, Huazhong Agricultural University , Wuhan 430070, China.

Both recent studies and our calculation suggest that the physicochemical

properties of launched drugs changed continuously over the past decades. Besides

shifting of commonly used properties, the average biological relevance (BR) and

similarity to natural products (NPs) of launched drugs decreased, reflecting the

fact that current drug discovery deviated away from NPs. To change the current

situation characterized by high investment but low productivity in drug

discovery, efforts should be made to improve the BR of the screening library and

hunt drugs more effectively in the biologically relevant chemical space.

Additionally, a multiple dimensional molecular descriptor, named the biologically

relevant spectrum (BRS) was proposed for quantitative structure-activity

relationships (QSAR) study or screening library preparation. Prediction models

for 43 biological activity categories were developed with BRS and support vector

machine (SVM). In most cases, the overall prediction accuracies were around 95%

and the Matthew's correlation coefficients (MCC) were over 0.8. Thirty-seven out

of 48 drug-activity associations were successfully predicted for drugs that

launched from 2006 to 2012, which were not included in the training data set. A

web-server named BioRel ( http://ibi.hzau.edu.cn/biorel ) was developed to

provide services including BR, BRS calculation, activity class, and

pharmacokinetic property prediction.

DOI: 10.1021/ci400432a

PMID: 24125686 [Indexed for MEDLINE]

985. J Theor Biol. 2013 Nov 21;337:71-9. doi: 10.1016/j.jtbi.2013.08.013. Epub 2013

Aug 26.

iCDI-PseFpt: identify the channel-drug interaction in cellular networking with

PseAAC and molecular fingerprints.

Xiao X(1), Min JL(2), Wang P(3), Chou KC(4).

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China; Information School, Zhe-Jiang Textile & Fashion College, Ning-Bo 315211,

China; Gordon Life Science Institute, 53 South Cottage Road, Belmont, MA 02478,

United States. Electronic address: xxiao@gordonlifescience.org. (2)Computer

Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China.

Electronic address: minjianliang@aliyun.com. (3)Computer Department, Jing-De-Zhen

Ceramic Institute, Jing-De-Zhen 333403, China. Electronic address:

wp3751@163.com. (4)Center of Excellence in Genomic Medicine Research (CEGMR),

King Abdulaziz University, Jeddah, Saudi Arabia; Gordon Life Science Institute,

53 South Cottage Road, Belmont, MA 02478, United States. Electronic address:

kcchou@gordonlifescience.org.

Many crucial functions in life, such as heartbeat, sensory transduction and

central nervous system response, are controlled by cell signalings via various

ion channels. Therefore, ion channels have become an excellent drug target, and

study of ion channel-drug interaction networks is an important topic for drug

development. However, it is both time-consuming and costly to determine whether a

drug and a protein ion channel are interacting with each other in a cellular

network by means of experimental techniques. Although some computational methods

were developed in this regard based on the knowledge of the 3D

(three-dimensional) structure of protein, unfortunately their usage is quite

limited because the 3D structures for most protein ion channels are still

unknown. With the avalanche of protein sequences generated in the post-genomic

age, it is highly desirable to develop the sequence-based computational method to

address this problem. To take up the challenge, we developed a new predictor

called iCDI-PseFpt, in which the protein ion-channel sample is formulated by the

PseAAC (pseudo amino acid composition) generated with the gray model theory, the

drug compound by the 2D molecular fingerprint, and the operation engine is the

fuzzy K-nearest neighbor algorithm. The overall success rate achieved by

iCDI-PseFpt via the jackknife cross-validation was 87.27%, which is remarkably

higher than that by any of the existing predictors in this area. As a

user-friendly web-server, iCDI-PseFpt is freely accessible to the public at the

website http://www.jci-bioinfo.cn/iCDI-PseFpt/. Furthermore, for the convenience

of most experimental scientists, a step-by-step guide is provided on how to use

the web-server to get the desired results without the need to follow the

complicated math equations presented in the paper just for its integrity. It has

not escaped our notice that the current approach can also be used to study other

drug-target interaction networks.

© 2013 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2013.08.013

PMID: 23988798 [Indexed for MEDLINE]

986. PLoS One. 2013 Nov 20;8(11):e79480. doi: 10.1371/journal.pone.0079480.

eCollection 2013.

Cube - an online tool for comparison and contrasting of protein sequences.

Zhang ZH(1), Khoo AA(1), Mihalek I(1).

Author information:

(1)Bioinformatics Institute, Agency for Science, Technology and Research,

Singapore.

When comparing sequences of similar proteins, two kinds of questions can be

asked, and the related two kinds of inference made. First, one may ask to what

degree they are similar, and then, how they differ. In the first case one may

tentatively conclude that the conserved elements common to all sequences are of

central and common importance to the protein's function. In the latter case the

regions of specialization may be discriminative of the function or binding

partners across subfamilies of related proteins. Experimental efforts -

mutagenesis or pharmacological intervention - can then be pointed in either

direction, depending on the context of the study. Cube simplifies this process

for users that already have their favorite sets of sequences, and helps them

collate the information by visualization of the conservation and specialization

scores on the sequence and on the structure, and by spreadsheet tabulation. All

information can be visualized on the spot, or downloaded for reference and later

inspection.SERVER HOMEPAGE: http://eopsf.org/cube.

DOI: 10.1371/journal.pone.0079480

PMCID: PMC3867285

PMID: 24363790 [Indexed for MEDLINE]

987. J Cheminform. 2013 Nov 18;5(1):46. doi: 10.1186/1758-2946-5-46.

CH5M3D: an HTML5 program for creating 3D molecular structures.

Earley CW(1).

Author information:

(1)Department of Chemistry, Kent State University at Stark, 6000 Frank Ave, NW,

North Canton, OH 44720, USA. cearley@kent.edu.

BACKGROUND: While a number of programs and web-based applications are available

for the interactive display of 3-dimensional molecular structures, few of these

provide the ability to edit these structures. For this reason, we have developed

a library written in JavaScript to allow for the simple creation of web-based

applications that should run on any browser capable of rendering HTML5 web pages.

While our primary interest in developing this application was for educational

use, it may also prove useful to researchers who want a light-weight application

for viewing and editing small molecular structures.

RESULTS: Molecular compounds are drawn on the HTML5 Canvas element, with the

JavaScript code making use of standard techniques to allow display of

three-dimensional structures on a two-dimensional canvas. Information about the

structure (bond lengths, bond angles, and dihedral angles) can be obtained using

a mouse or other pointing device. Both atoms and bonds can be added or deleted,

and rotation about bonds is allowed. Routines are provided to read structures

either from the web server or from the user's computer, and creation of galleries

of structures can be accomplished with only a few lines of code. Documentation

and examples are provided to demonstrate how users can access all of the

molecular information for creation of web pages with more advanced features.

CONCLUSIONS: A light-weight (≈ 75 kb) JavaScript library has been made available

that allows for the simple creation of web pages containing interactive

3-dimensional molecular structures. Although this library is designed to create

web pages, a web server is not required. Installation on a web server is

straightforward and does not require any server-side modules or special

permissions. The ch5m3d.js library has been released under the GNU GPL version 3

open-source license and is available from

http://sourceforge.net/projects/ch5m3d/.

DOI: 10.1186/1758-2946-5-46

PMCID: PMC4177146

PMID: 24246004

988. Bioinformatics. 2013 Nov 15;29(22):2852-8. doi: 10.1093/bioinformatics/btt506.

Epub 2013 Sep 4.

INSECT: IN-silico SEarch for Co-occurring Transcription factors.

Rohr CO(1), Parra RG, Yankilevich P, Perez-Castro C.

Author information:

(1)Instituto de Ecología, Genética y Evolución (IEGEBA)-CONICET, Protein

Physiology Laboratory, Departamento de Química Biológica, Facultad de Ciencias

Exactas y Naturales, UBA-CONICET-IQUIBICEN, C1428EGA, Buenos Aires, Argentina and

Instituto de Investigación en Biomedicina de Buenos Aires (IBioBA) -CONICET-

Partner Institute of the Max Planck Society, C1425FQD, Buenos Aires, Argentina.

MOTIVATION: Transcriptional regulation occurs through the concerted actions of

multiple transcription factors (TFs) that bind cooperatively to cis-regulatory

modules (CRMs) of genes. These CRMs usually contain a variable number of

transcription factor-binding sites (TFBSs) involved in related cellular and

physiological processes. Chromatin immunoprecipitation followed by sequencing

(ChIP-seq) has been effective in detecting TFBSs and nucleosome location to

identify potential CRMs in genome-wide studies. Although several attempts were

previously reported to predict the potential binding of TFs at TFBSs within CRMs

by comparing different ChIP-seq data, these have been hampered by excessive

background, usually emerging as a consequence of experimental conditions. To

understand these complex regulatory circuits, it would be helpful to have

reliable and updated user-friendly tools to assist in the identification of TFBSs

and CRMs for gene(s) of interest.

RESULTS: Here we present INSECT (IN-silico SEarch for Co-occurring Transcription

factors), a novel web server for identifying potential TFBSs and CRMs in gene

sequences. By combining several strategies, INSECT provides flexible analysis of

multiple co-occurring TFBSs, by applying differing search schemes and restriction

parameters. availability and implementation: INSECT is freely available as a web

server at http://bioinformatics.ibioba-mpsp-conicet.gov.ar/INSECT .

DOI: 10.1093/bioinformatics/btt506

PMID: 24008418 [Indexed for MEDLINE]

989. Bioinformatics. 2013 Nov 15;29(22):2953-4. doi: 10.1093/bioinformatics/btt507.

Epub 2013 Sep 3.

M2SG: mapping human disease-related genetic variants to protein sequences and

genomic loci.

Ji R(1), Cong Q, Li W, Grishin NV.

Author information:

(1)Departments of biophysics and biochemistry, University of Texas Southwestern

Medical Center and Howard Hughes Medical Institute, University of Texas

Southwestern Medical Center, Dallas, TX 75390-9050, USA.

SUMMARY: Online Mendelian Inheritance in Man (OMIM) is a manually curated

compendium of human genetic variants and the corresponding phenotypes, mostly

human diseases. Instead of directly documenting the native sequences for gene

entries, OMIM links its entries to protein and DNA sequences in other databases.

However, because of the existence of gene isoforms and errors in OMIM records,

mapping a specific OMIM mutation to its corresponding protein sequence is not

trivial. Combining computer programs and extensive manual curation of OMIM

full-text descriptions and original literature, we mapped 98% of OMIM amino acid

substitutions (AASs) and all SwissProt Variant (SwissVar) disease-related AASs to

reference sequences and confidently mapped 99.96% of all AASs to the genomic

loci. Based on the results, we developed an online database and interactive web

server (M2SG) to (i) retrieve the mapped OMIM and SwissVar variants for a given

protein sequence; and (ii) obtain related proteins and mutations for an input

disease phenotype. This database will be useful for analyzing sequences,

understanding the effect of mutations, identifying important genetic variations

and designing experiments on a protein of interest.

AVAILABILITY AND IMPLEMENTATION: The database and web server are freely available

at http://prodata.swmed.edu/M2S/mut2seq.cgi.

DOI: 10.1093/bioinformatics/btt507

PMCID: PMC3810852

PMID: 24002112 [Indexed for MEDLINE]

990. Bioinformatics. 2013 Nov 15;29(22):2931-2. doi: 10.1093/bioinformatics/btt501.

Epub 2013 Aug 30.

TALENoffer: genome-wide TALEN off-target prediction.

Grau J(1), Boch J, Posch S.

Author information:

(1)Institute of Computer Science and Department of Genetics, Institute of

Biology, Martin Luther University Halle-Wittenberg, D-06099 Halle (Saale),

Germany.

SUMMARY: Transcription activator-like effector nucleases (TALENs) have become an

accepted tool for targeted mutagenesis, but undesired off-targets remain an

important issue. We present TALENoffer, a novel tool for the genome-wide

prediction of TALEN off-targets. We show that TALENoffer successfully predicts

known off-targets of engineered TALENs and yields a competitive runtime, scanning

complete mammalian genomes within a few minutes.

AVAILABILITY: TALENoffer is available as a command line program from

http://www.jstacs.de/index.php/TALENoffer and as a Galaxy server at

http://galaxy.informatik.uni-halle.de.

CONTACT: grau@informatik.uni-halle.de

DOI: 10.1093/bioinformatics/btt501

PMID: 23995255 [Indexed for MEDLINE]

991. Bioinformatics. 2013 Nov 15;29(22):2928-30. doi: 10.1093/bioinformatics/btt495.

Epub 2013 Aug 23.

catRAPID omics: a web server for large-scale prediction of protein-RNA

interactions.

Agostini F(1), Zanzoni A, Klus P, Marchese D, Cirillo D, Tartaglia GG.

Author information:

(1)Gene Function and Evolution, Bioinformatics and Genomics, Centre for Genomic

Regulation (CRG) and Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.

SUMMARY: Here we introduce catRAPID omics, a server for large-scale calculations

of protein-RNA interactions. Our web server allows (i) predictions at proteomic

and transcriptomic level; (ii) use of protein and RNA sequences without size

restriction; (iii) analysis of nucleic acid binding regions in proteins; and (iv)

detection of RNA motifs involved in protein recognition.

RESULTS: We developed a web server to allow fast calculation of ribonucleoprotein

associations in Caenorhabditis elegans, Danio rerio, Drosophila melanogaster,

Homo sapiens, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae and

Xenopus tropicalis (custom libraries can be also generated). The catRAPID omics

was benchmarked on the recently published RNA interactomes of

Serine/arginine-rich splicing factor 1 (SRSF1), Histone-lysine

N-methyltransferase EZH2 (EZH2), TAR DNA-binding protein 43 (TDP43) and

RNA-binding protein FUS (FUS) as well as on the protein interactomes of U1/U2

small nucleolar RNAs, X inactive specific transcript (Xist) repeat A region

(RepA) and Crumbs homolog 3 (CRB3) 3'-untranslated region RNAs. Our predictions

are highly significant (P < 0.05) and will help the experimentalist to identify

candidates for further validation.

AVAILABILITY: catRAPID omics can be freely accessed on the Web at

http://s.tartaglialab.com/catrapid/omics. Documentation, tutorial and FAQs are

available at http://s.tartaglialab.com/page/catrapid\_group.

DOI: 10.1093/bioinformatics/btt495

PMCID: PMC3810848

PMID: 23975767 [Indexed for MEDLINE]

992. PLoS One. 2013 Nov 14;8(11):e80170. doi: 10.1371/journal.pone.0080170.

eCollection 2013.

CNVannotator: a comprehensive annotation server for copy number variation in the

human genome.

Zhao M(1), Zhao Z.

Author information:

(1)Department of Biomedical Informatics, Vanderbilt University School of

Medicine, Nashville, Tennessee, United States of America.

Copy number variation (CNV) is one of the most prevalent genetic variations in

the genome, leading to an abnormal number of copies of moderate to large genomic

regions. High-throughput technologies such as next-generation sequencing often

identify thousands of CNVs involved in biological or pathological processes.

Despite the growing demand to filter and classify CNVs by factors such as

frequency in population, biological features, and function, surprisingly, no

online web server for CNV annotations has been made available to the research

community. Here, we present CNVannotator, a web server that accepts an input set

of human genomic positions in a user-friendly tabular format. CNVannotator can

perform genomic overlaps of the input coordinates using various functional

features, including a list of the reported 356,817 common CNVs, 181,261 disease

CNVs, as well as, 140,342 SNPs from genome-wide association studies. In addition,

CNVannotator incorporates 2,211,468 genomic features, including ENCODE regulatory

elements, cytoband, segmental duplication, genome fragile site, pseudogene,

promoter, enhancer, CpG island, and methylation site. For cancer research

community users, CNVannotator can apply various filters to retrieve a subgroup of

CNVs pinpointed in hundreds of tumor suppressor genes and oncogenes. In total,

5,277,234 unique genomic coordinates with functional features are available to

generate an output in a plain text format that is free to download. In summary,

we provide a comprehensive web resource for human CNVs. The annotated results

along with the server can be accessed at

http://bioinfo.mc.vanderbilt.edu/CNVannotator/.

DOI: 10.1371/journal.pone.0080170

PMCID: PMC3828214

PMID: 24244640 [Indexed for MEDLINE]

993. PLoS One. 2013 Nov 12;8(11):e79489. doi: 10.1371/journal.pone.0079489.

eCollection 2013.

Using a web-based application to define the accuracy of diagnostic tests when the

gold standard is imperfect.

Lim C(1), Wannapinij P, White L, Day NP, Cooper BS, Peacock SJ, Limmathurotsakul

D.

Author information:

(1)Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine,

Mahidol University, Bangkok, Thailand.

BACKGROUND: Estimates of the sensitivity and specificity for new diagnostic tests

based on evaluation against a known gold standard are imprecise when the accuracy

of the gold standard is imperfect. Bayesian latent class models (LCMs) can be

helpful under these circumstances, but the necessary analysis requires expertise

in computational programming. Here, we describe open-access web-based

applications that allow non-experts to apply Bayesian LCMs to their own data sets

via a user-friendly interface.

METHODS/PRINCIPAL FINDINGS: Applications for Bayesian LCMs were constructed on a

web server using R and WinBUGS programs. The models provided

(http://mice.tropmedres.ac) include two Bayesian LCMs: the two-tests in

two-population model (Hui and Walter model) and the three-tests in one-population

model (Walter and Irwig model). Both models are available with simplified and

advanced interfaces. In the former, all settings for Bayesian statistics are

fixed as defaults. Users input their data set into a table provided on the

webpage. Disease prevalence and accuracy of diagnostic tests are then estimated

using the Bayesian LCM, and provided on the web page within a few minutes. With

the advanced interfaces, experienced researchers can modify all settings in the

models as needed. These settings include correlation among diagnostic test

results and prior distributions for all unknown parameters. The web pages provide

worked examples with both models using the original data sets presented by Hui

and Walter in 1980, and by Walter and Irwig in 1988. We also illustrate the

utility of the advanced interface using the Walter and Irwig model on a data set

from a recent melioidosis study. The results obtained from the web-based

applications were comparable to those published previously.

CONCLUSIONS: The newly developed web-based applications are open-access and

provide an important new resource for researchers worldwide to evaluate new

diagnostic tests.

DOI: 10.1371/journal.pone.0079489

PMCID: PMC3827152

PMID: 24265775 [Indexed for MEDLINE]

994. PLoS One. 2013 Nov 12;8(11):e78383. doi: 10.1371/journal.pone.0078383.

eCollection 2013.

PANADA: protein association network annotation, determination and analysis.

Martin AJ(1), Walsh I, Domenico TD, Mičetić I, Tosatto SC.

Author information:

(1)Department of Biology, University of Padova, Padova, Italy.

Increasingly large numbers of proteins require methods for functional annotation.

This is typically based on pairwise inference from the homology of either protein

sequence or structure. Recently, similarity networks have been presented to

leverage both the ability to visualize relationships between proteins and assess

the transferability of functional inference. Here we present PANADA, a novel

toolkit for the visualization and analysis of protein similarity networks in

Cytoscape. Networks can be constructed based on pairwise sequence or structural

alignments either on a set of proteins or, alternatively, by database search from

a single sequence. The Panada web server, executable for download and examples

and extensive help files are available at URL:

http://protein.bio.unipd.it/panada/.

DOI: 10.1371/journal.pone.0078383

PMCID: PMC3827049

PMID: 24265686 [Indexed for MEDLINE]

995. BMC Genomics. 2013 Nov 10;14:774. doi: 10.1186/1471-2164-14-774.

BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data.

Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, Chen PY(1), Pellegrini M.

Author information:

(1)Department of Molecular, Cell and Developmental Biology, University of

California, Los Angeles, CA 90095, USA. paoyang@gate.sinica.edu.tw.

BACKGROUND: DNA methylation is an important epigenetic modification involved in

many biological processes. Bisulfite treatment coupled with high-throughput

sequencing provides an effective approach for studying genome-wide DNA

methylation at base resolution. Libraries such as whole genome bisulfite

sequencing (WGBS) and reduced represented bisulfite sequencing (RRBS) are widely

used for generating DNA methylomes, demanding efficient and versatile tools for

aligning bisulfite sequencing data.

RESULTS: We have developed BS-Seeker2, an updated version of BS Seeker, as a full

pipeline for mapping bisulfite sequencing data and generating DNA methylomes.

BS-Seeker2 improves mappability over existing aligners by using local alignment.

It can also map reads from RRBS library by building special indexes with improved

efficiency and accuracy. Moreover, BS-Seeker2 provides additional function for

filtering out reads with incomplete bisulfite conversion, which is useful in

minimizing the overestimation of DNA methylation levels. We also defined CGmap

and ATCGmap file formats for full representations of DNA methylomes, as part of

the outputs of BS-Seeker2 pipeline together with BAM and WIG files.

CONCLUSIONS: Our evaluations on the performance show that BS-Seeker2 works

efficiently and accurately for both WGBS data and RRBS data. BS-Seeker2 is freely

available at http://pellegrini.mcdb.ucla.edu/BS\_Seeker2/ and the Galaxy server.

DOI: 10.1186/1471-2164-14-774

PMCID: PMC3840619

PMID: 24206606 [Indexed for MEDLINE]

996. J Theor Biol. 2013 Nov 7;336:11-7. doi: 10.1016/j.jtbi.2013.07.009. Epub 2013 Jul

18.

Prediction of pupylation sites using the composition of k-spaced amino acid

pairs.

Tung CW(1).

Author information:

(1)School of Pharmacy, Kaohsiung Medical University, Kaohsiung 807, Taiwan; PhD

Program in Toxicology, Kaohsiung Medical University, Kaohsiung 807, Taiwan.

Electronic address: cwtung@kmu.edu.tw.

Pupylation is an important post-translational modification in prokaryotes. A

prokaryotic ubiquitin-like protein (Pup) is attached to proteins as a signal for

selective degradation by proteasome. Several proteomics methods have been

developed for the identification of pupylated proteins and pupylation sites.

However, pupylation sites of many experimentally identified pupylated proteins

are still unknown. The development of sequence-based prediction methods can help

to accelerate the identification of pupylation sites and gain insights into the

substrate specificity and regulatory functions of pupylation. A novel tool iPUP

is developed for the computational identification of pupylation sites. A

composition of k-spaced amino acid pairs is utilized to represent a peptide

sequence. Top ranked k-spaced amino acid pairs are subsequently selected by using

a sequential backward feature elimination algorithm. The 10-fold cross-validation

performance of iPUP trained by using the composition of 150 top ranked k-spaced

amino acid pairs and support vector machines is 0.83 for the area under receiver

operating characteristic curve. The importance analysis of k-spaced amino acid

pairs shows that terminal space-containing pairs are useful for discriminating

pupylation sites from non-pupylation sites. A sequence analysis confirms that

lysines close to C-terminus tend to be pupylated. In contrast, lysines close to

N-terminus are less likely to be pupylated. The iPUP tool can predict pupylation

sites with probability scores for prioritizing promising pupylation sites. Both

the online server and the standalone software of iPUP are freely available for

academic use at http://cwtung.kmu.edu.tw/ipup.

Copyright © 2013 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2013.07.009

PMID: 23871866 [Indexed for MEDLINE]

997. PLoS One. 2013 Nov 5;8(11):e79288. doi: 10.1371/journal.pone.0079288. eCollection

2013.

Viral IRES prediction system - a web server for prediction of the IRES secondary

structure in silico.

Hong JJ(1), Wu TY, Chang TY, Chen CY.

Author information:

(1)Department of Bioscience Technology, Chung Yuan Christian University,

Chung-Li, Taiwan.

The internal ribosomal entry site (IRES) functions as cap-independent translation

initiation sites in eukaryotic cells. IRES elements have been applied as useful

tools for bi-cistronic expression vectors. Current RNA structure prediction

programs are unable to predict precisely the potential IRES element. We have

designed a viral IRES prediction system (VIPS) to perform the IRES secondary

structure prediction. In order to obtain better results for the IRES prediction,

the VIPS can evaluate and predict for all four different groups of IRESs with a

higher accuracy. RNA secondary structure prediction, comparison, and pseudoknot

prediction programs were implemented to form the three-stage procedure for the

VIPS. The backbone of VIPS includes: the RNAL fold program, aimed to predict

local RNA secondary structures by minimum free energy method; the RNA Align

program, intended to compare predicted structures; and pknotsRG program, used to

calculate the pseudoknot structure. VIPS was evaluated by using UTR database,

IRES database and Virus database, and the accuracy rate of VIPS was assessed as

98.53%, 90.80%, 82.36% and 80.41% for IRES groups 1, 2, 3, and 4, respectively.

This advance useful search approach for IRES structures will facilitate IRES

related studies. The VIPS on-line website service is available at

http://140.135.61.250/vips/.

DOI: 10.1371/journal.pone.0079288

PMCID: PMC3818432

PMID: 24223923 [Indexed for MEDLINE]

998. Acta Crystallogr D Biol Crystallogr. 2013 Nov;69(Pt 11):2167-73. doi:

10.1107/S0907444913015291. Epub 2013 Oct 12.

Molecular replacement: tricks and treats.

Abergel C(1).

Author information:

(1)Information Génomique et Structurale, IGS UMR 7256, CNRS, Aix-Marseille

Université, IMM, FR3479, 163 Avenue de Luminy - case 934, 13288 Marseille CEDEX

09, France.

Molecular replacement is the method of choice for X-ray crystallographic

structure determination provided that suitable structural homologues are

available in the PDB. Presently, there are ~80,000 structures in the PDB (8074

were deposited in the year 2012 alone), of which ~70% have been solved by

molecular replacement. For successful molecular replacement the model must cover

at least 50% of the total structure and the Cα r.m.s.d. between the core model

and the structure to be solved must be less than 2 Å. Here, an approach

originally implemented in the CaspR server

(http://www.igs.cnrs-mrs.fr/Caspr2/index.cgi) based on homology modelling to

search for a molecular-replacement solution is discussed. How the use of as much

information as possible from different sources can improve the model(s) is

briefly described. The combination of structural information with distantly

related sequences is crucial to optimize the multiple alignment that will define

the boundaries of the core domains. PDB clusters (sequences with ≥30% identical

residues) can also provide information on the eventual changes in conformation

and will help to explore the relative orientations assumed by protein subdomains.

Normal-mode analysis can also help in generating series of conformational models

in the search for a molecular-replacement solution. Of course, finding a correct

solution is only the first step and the accuracy of the identified solution is as

important as the data quality to proceed through refinement. Here, some possible

reasons for failure are discussed and solutions are proposed using a set of

successful examples.

DOI: 10.1107/S0907444913015291

PMCID: PMC3817689

PMID: 24189227 [Indexed for MEDLINE]

999. Aliment Pharmacol Ther. 2013 Nov;38(9):1109-18. doi: 10.1111/apt.12493. Epub 2013

Sep 17.

The management of iron deficiency in inflammatory bowel disease--an online tool

developed by the RAND/UCLA appropriateness method.

Reinisch W(1), Chowers Y, Danese S, Dignass A, Gomollón F, Nielsen OH, Lakatos

PL, Lees CW, Lindgren S, Lukas M, Mantzaris GJ, Michetti P, Moum B,

Peyrin-Biroulet L, Toruner M, van der Woude J, Weiss G, Stoevelaar H.

Author information:

(1)Department Internal Medicine III, Medical University of Vienna, Vienna,

Austria.

BACKGROUND: Iron deficiency is a common and undertreated problem in inflammatory

bowel disease (IBD).

AIM: To develop an online tool to support treatment choice at the

patient-specific level.

METHODS: Using the RAND/UCLA Appropriateness Method (RUAM), a European expert

panel assessed the appropriateness of treatment regimens for a variety of

clinical scenarios in patients with non-anaemic iron deficiency (NAID) and iron

deficiency anaemia (IDA). Treatment options included adjustment of IBD medication

only, oral iron supplementation, high-/low-dose intravenous (IV) regimens, IV

iron plus erythropoietin-stimulating agent (ESA), and blood transfusion. The

panel process consisted of two individual rating rounds (1148 treatment

indications; 9-point scale) and three plenary discussion meetings.

RESULTS: The panel reached agreement on 71% of treatment indications. 'No

treatment' was never considered appropriate, and repeat treatment after previous

failure was generally discouraged. For 98% of scenarios, at least one treatment

was appropriate. Adjustment of IBD medication was deemed appropriate in all

patients with active disease. Use of oral iron was mainly considered an option in

NAID and mildly anaemic patients without disease activity. IV regimens were often

judged appropriate, with high-dose IV iron being the preferred option in 77% of

IDA scenarios. Blood transfusion and IV+ESA were indicated in exceptional cases

only.

CONCLUSIONS: The RUAM revealed high agreement amongst experts on the management

of iron deficiency in patients with IBD. High-dose IV iron was more often

considered appropriate than other options. To facilitate dissemination of the

recommendations, panel outcomes were embedded in an online tool, accessible via

http://ferroscope.com/.

© 2013 John Wiley & Sons Ltd.

DOI: 10.1111/apt.12493

PMCID: PMC3886296

PMID: 24099472 [Indexed for MEDLINE]

1000. Anal Biochem. 2013 Nov 1;442(1):118-25. doi: 10.1016/j.ab.2013.05.024. Epub 2013

Jun 10.

iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced

amino acid alphabet composition.

Feng PM(1), Chen W, Lin H, Chou KC.

Author information:

(1)School of Public Health, Hebei United University, Tangshan 063000, China.

Heat shock proteins (HSPs) are a type of functionally related proteins present in

all living organisms, both prokaryotes and eukaryotes. They play essential roles

in protein-protein interactions such as folding and assisting in the

establishment of proper protein conformation and prevention of unwanted protein

aggregation. Their dysfunction may cause various life-threatening disorders, such

as Parkinson's, Alzheimer's, and cardiovascular diseases. Based on their

functions, HSPs are usually classified into six families: (i) HSP20 or sHSP, (ii)

HSP40 or J-class proteins, (iii) HSP60 or GroEL/ES, (iv) HSP70, (v) HSP90, and

(vi) HSP100. Although considerable progress has been achieved in discriminating

HSPs from other proteins, it is still a big challenge to identify HSPs among

their six different functional types according to their sequence information

alone. With the avalanche of protein sequences generated in the post-genomic age,

it is highly desirable to develop a high-throughput computational tool in this

regard. To take up such a challenge, a predictor called iHSP-PseRAAAC has been

developed by incorporating the reduced amino acid alphabet information into the

general form of pseudo amino acid composition. One of the remarkable advantages

of introducing the reduced amino acid alphabet is being able to avoid the

notorious dimension disaster or overfitting problem in statistical prediction. It

was observed that the overall success rate achieved by iHSP-PseRAAAC in

identifying the functional types of HSPs among the aforementioned six types was

more than 87%, which was derived by the jackknife test on a stringent benchmark

dataset in which none of HSPs included has ≥40% pairwise sequence identity to any

other in the same subset. It has not escaped our notice that the reduced amino

acid alphabet approach can also be used to investigate other protein

classification problems. As a user-friendly web server, iHSP-PseRAAAC is

accessible to the public at http://lin.uestc.edu.cn/server/iHSP-PseRAAAC.

Copyright © 2013 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2013.05.024

PMID: 23756733 [Indexed for MEDLINE]

1001. Bioinformatics. 2013 Nov 1;29(21):2722-8. doi: 10.1093/bioinformatics/btt473.

Epub 2013 Aug 27.

lDDT: a local superposition-free score for comparing protein structures and

models using distance difference tests.

Mariani V(1), Biasini M, Barbato A, Schwede T.

Author information:

(1)Biozentrum, Universität Basel, Klingelbergstrasse 50-70 and Computational

Structural Biology, SIB Swiss Institute of Bioinformatics, 4056 Basel,

Switzerland.

MOTIVATION: The assessment of protein structure prediction techniques requires

objective criteria to measure the similarity between a computational model and

the experimentally determined reference structure. Conventional similarity

measures based on a global superposition of carbon α atoms are strongly

influenced by domain motions and do not assess the accuracy of local atomic

details in the model.

RESULTS: The Local Distance Difference Test (lDDT) is a superposition-free score

that evaluates local distance differences of all atoms in a model, including

validation of stereochemical plausibility. The reference can be a single

structure, or an ensemble of equivalent structures. We demonstrate that lDDT is

well suited to assess local model quality, even in the presence of domain

movements, while maintaining good correlation with global measures. These

properties make lDDT a robust tool for the automated assessment of structure

prediction servers without manual intervention.

AVAILABILITY AND IMPLEMENTATION: Source code, binaries for Linux and MacOSX, and

an interactive web server are available at http://swissmodel.expasy.org/lddt.

CONTACT: torsten.schwede@unibas.ch.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt473

PMCID: PMC3799472

PMID: 23986568 [Indexed for MEDLINE]

1002. Mol Phylogenet Evol. 2013 Nov;69(2):313-9. doi: 10.1016/j.ympev.2012.08.023. Epub

2012 Sep 7.

MITOS: improved de novo metazoan mitochondrial genome annotation.

Bernt M(1), Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J,

Middendorf M, Stadler PF.

Author information:

(1)Parallel Computing and Complex Systems Group, Department of Computer Science,

University Leipzig, Augustusplatz 10-11, 04109 Leipzig, Germany. Electronic

address: bernt@informatik.uni-leipzig.de.

About 2000 completely sequenced mitochondrial genomes are available from the NCBI

RefSeq data base together with manually curated annotations of their

protein-coding genes, rRNAs, and tRNAs. This annotation information, which has

accumulated over two decades, has been obtained with a diverse set of

computational tools and annotation strategies. Despite all efforts of manual

curation it is still plagued by misassignments of reading directions, erroneous

gene names, and missing as well as false positive annotations in particular for

the RNA genes. Taken together, this causes substantial problems for fully

automatic pipelines that aim to use these data comprehensively for studies of

animal phylogenetics and the molecular evolution of mitogenomes. The MITOS

pipeline is designed to compute a consistent de novo annotation of the

mitogenomic sequences. We show that the results of MITOS match RefSeq and MitoZoa

in terms of annotation coverage and quality. At the same time we avoid biases,

inconsistencies of nomenclature, and typos originating from manual curation

strategies. The MITOS pipeline is accessible online at

http://mitos.bioinf.uni-leipzig.de.

Copyright © 2012 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ympev.2012.08.023

PMID: 22982435 [Indexed for MEDLINE]

1003. Biol Direct. 2013 Oct 30;8:27. doi: 10.1186/1745-6150-8-27.

Identification of B-cell epitopes in an antigen for inducing specific class of

antibodies.

Gupta S, Ansari HR, Gautam A; Open Source Drug Discovery Consortium, Raghava

GP(1).

Author information:

(1)Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh

160036, India. raghava@imtech.res.in.

BACKGROUND: In the past, numerous methods have been developed for predicting

antigenic regions or B-cell epitopes that can induce B-cell response. To the best

of authors' knowledge, no method has been developed for predicting B-cell

epitopes that can induce a specific class of antibody (e.g., IgA, IgG) except

allergenic epitopes (IgE). In this study, an attempt has been made to understand

the relation between primary sequence of epitopes and the class of antibodies

generated.

RESULTS: The dataset used in this study has been derived from Immune Epitope

Database and consists of 14725 B-cell epitopes that include 11981 IgG, 2341 IgE,

403 IgA specific epitopes and 22835 non-B-cell epitopes. In order to understand

the preference of residues or motifs in these epitopes, we computed and compared

amino acid and dipeptide composition of IgG, IgE, IgA inducing epitopes and

non-B-cell epitopes. Differences in composition profiles of different classes of

epitopes were observed, and few residues were found to be preferred. Based on

these observations, we developed models for predicting antibody class-specific

B-cell epitopes using various features like amino acid composition, dipeptide

composition, and binary profiles. Among these, dipeptide composition-based

support vector machine model achieved maximum Matthews correlation coefficient of

0.44, 0.70 and 0.45 for IgG, IgE and IgA specific epitopes respectively. All

models were developed on experimentally validated non-redundant dataset and

evaluated using five-fold cross validation. In addition, the performance of

dipeptide-based model was also evaluated on independent dataset.

CONCLUSION: Present study utilizes the amino acid sequence information for

predicting the tendencies of antigens to induce different classes of antibodies.

For the first time, in silico models have been developed for predicting B-cell

epitopes, which can induce specific class of antibodies. A web service called

IgPred has been developed to serve the scientific community. This server will be

useful for researchers working in the field of subunit/epitope/peptide-based

vaccines and immunotherapy (http://crdd.osdd.net/raghava/igpred/).

DOI: 10.1186/1745-6150-8-27

PMCID: PMC3831251

PMID: 24168386 [Indexed for MEDLINE]

1004. PLoS One. 2013 Oct 29;8(10):e76864. doi: 10.1371/journal.pone.0076864.

eCollection 2013.

SeqNLS: nuclear localization signal prediction based on frequent pattern mining

and linear motif scoring.

Lin JR(1), Hu J.

Author information:

(1)Department of Computer Science and Engineering, University of South Carolina,

Columbia, South Carolina, United States of America.

Nuclear localization signals (NLSs) are stretches of residues in proteins

mediating their importing into the nucleus. NLSs are known to have diverse

patterns, of which only a limited number are covered by currently known NLS

motifs. Here we propose a sequential pattern mining algorithm SeqNLS to

effectively identify potential NLS patterns without being constrained by the

limitation of current knowledge of NLSs. The extracted frequent sequential

patterns are used to predict NLS candidates which are then filtered by a linear

motif-scoring scheme based on predicted sequence disorder and by the relatively

local conservation (IRLC) based masking. The experiment results on the newly

curated Yeast and Hybrid datasets show that SeqNLS is effective in detecting

potential NLSs. The performance comparison between SeqNLS with and without the

linear motif scoring shows that linear motif features are highly complementary to

sequence features in discerning NLSs. For the two independent datasets, our

SeqNLS not only can consistently find over 50% of NLSs with prediction precision

of at least 0.7, but also outperforms other state-of-the-art NLS prediction

methods in terms of F1 score or prediction precision with similar or higher

recall rates. The web server of the SeqNLS algorithm is available at

http://mleg.cse.sc.edu/seqNLS.

DOI: 10.1371/journal.pone.0076864

PMCID: PMC3812174

PMID: 24204689 [Indexed for MEDLINE]

1005. PLoS One. 2013 Oct 28;8(10):e77302. doi: 10.1371/journal.pone.0077302.

eCollection 2013.

PathogenFinder--distinguishing friend from foe using bacterial whole genome

sequence data.

Cosentino S(1), Voldby Larsen M, Møller Aarestrup F, Lund O.

Author information:

(1)Center for Biological Sequence Analysis, Department of Systems Biology,

Technical University of Denmark, Kgs. Lyngby, Denmark.

Erratum in

PLoS One. 2013;8(12).

doi:10.1371/annotation/b84e1af7-c127-45c3-be22-76abd977600f.

Although the majority of bacteria are harmless or even beneficial to their host,

others are highly virulent and can cause serious diseases, and even death. Due to

the constantly decreasing cost of high-throughput sequencing there are now many

completely sequenced genomes available from both human pathogenic and innocuous

strains. The data can be used to identify gene families that correlate with

pathogenicity and to develop tools to predict the pathogenicity of newly

sequenced strains, investigations that previously were mainly done by means of

more expensive and time consuming experimental approaches. We describe

PathogenFinder (http://cge.cbs.dtu.dk/services/PathogenFinder/), a web-server for

the prediction of bacterial pathogenicity by analysing the input proteome,

genome, or raw reads provided by the user. The method relies on groups of

proteins, created without regard to their annotated function or known involvement

in pathogenicity. The method has been built to work with all taxonomic groups of

bacteria and using the entire training-set, achieved an accuracy of 88.6% on an

independent test-set, by correctly classifying 398 out of 449 completely

sequenced bacteria. The approach here proposed is not biased on sets of genes

known to be associated with pathogenicity, thus the approach could aid the

discovery of novel pathogenicity factors. Furthermore the pathogenicity

prediction web-server could be used to isolate the potential pathogenic features

of both known and unknown strains.

DOI: 10.1371/journal.pone.0077302

PMCID: PMC3810466

PMID: 24204795 [Indexed for MEDLINE]

1006. PLoS One. 2013 Oct 23;8(10):e77478. doi: 10.1371/journal.pone.0077478.

eCollection 2013.

Inference of gene-phenotype associations via protein-protein interaction and

orthology.

Wang P(1), Lai WF, Li MJ, Xu F, Yalamanchili HK, Lovell-Badge R, Wang J.

Author information:

(1)Department of Biochemistry, LKS Faculty of Medicine, The University of Hong

Kong, Hong Kong SAR, China ; Shenzhen Institute of Research and Innovation, The

University of Hong Kong, Shenzhen, China.

One of the fundamental goals of genetics is to understand gene functions and

their associated phenotypes. To achieve this goal, in this study we developed a

computational algorithm that uses orthology and protein-protein interaction

information to infer gene-phenotype associations for multiple species.

Furthermore, we developed a web server that provides genome-wide phenotype

inference for six species: fly, human, mouse, worm, yeast, and zebrafish. We

evaluated our inference method by comparing the inferred results with known

gene-phenotype associations. The high Area Under the Curve values suggest a

significant performance of our method. By applying our method to two human

representative diseases, Type 2 Diabetes and Breast Cancer, we demonstrated that

our method is able to identify related Gene Ontology terms and Kyoto Encyclopedia

of Genes and Genomes pathways. The web server can be used to infer functions and

putative phenotypes of a gene along with the candidate genes of a phenotype, and

thus aids in disease candidate gene discovery. Our web server is available at

http://jjwanglab.org/PhenoPPIOrth.

DOI: 10.1371/journal.pone.0077478

PMCID: PMC3806783

PMID: 24194887 [Indexed for MEDLINE]

1007. PLoS One. 2013 Oct 23;8(10):e77429. doi: 10.1371/journal.pone.0077429.

eCollection 2013.

CorSig: a general framework for estimating statistical significance of

correlation and its application to gene co-expression analysis.

Wang HQ(1), Tsai CJ.

Author information:

(1)Intelligent Computing Lab, Hefei Institutes of Physical Science, Chinese

Academy of Sciences, Hefei, China.

With the rapid increase of omics data, correlation analysis has become an

indispensable tool for inferring meaningful associations from a large number of

observations. Pearson correlation coefficient (PCC) and its variants are widely

used for such purposes. However, it remains challenging to test whether an

observed association is reliable both statistically and biologically. We present

here a new method, CorSig, for statistical inference of correlation significance.

CorSig is based on a biology-informed null hypothesis, i.e., testing whether the

true PCC (ρ) between two variables is statistically larger than a user-specified

PCC cutoff (τ), as opposed to the simple null hypothesis of ρ = 0 in existing

methods, i.e., testing whether an association can be declared without a

threshold. CorSig incorporates Fisher's Z transformation of the observed PCC (r),

which facilitates use of standard techniques for p-value computation and multiple

testing corrections. We compared CorSig against two methods: one uses a minimum

PCC cutoff while the other (Zhu's procedure) controls correlation strength and

statistical significance in two discrete steps. CorSig consistently outperformed

these methods in various simulation data scenarios by balancing between false

positives and false negatives. When tested on real-world Populus microarray data,

CorSig effectively identified co-expressed genes in the flavonoid pathway, and

discriminated between closely related gene family members for their differential

association with flavonoid and lignin pathways. The p-values obtained by CorSig

can be used as a stand-alone parameter for stratification of co-expressed genes

according to their correlation strength in lieu of an arbitrary cutoff. CorSig

requires one single tunable parameter, and can be readily extended to other

correlation measures. Thus, CorSig should be useful for a wide range of

applications, particularly for network analysis of high-dimensional genomic

data.SOFTWARE AVAILABILITY: A web server for CorSig is provided at

http://202.127.200.1:8080/probeWeb. R code for CorSig is freely available for

non-commercial use at http://aspendb.uga.edu/downloads.

DOI: 10.1371/journal.pone.0077429

PMCID: PMC3806744

PMID: 24194884 [Indexed for MEDLINE]

1008. PLoS One. 2013 Oct 23;8(10):e75826. doi: 10.1371/journal.pone.0075826.

eCollection 2013.

PSI: a comprehensive and integrative approach for accurate plant subcellular

localization prediction.

Liu L(1), Zhang Z, Mei Q, Chen M.

Author information:

(1)College of Life Sciences, Zhejiang University, Hangzhou, China.

Predicting the subcellular localization of proteins conquers the major drawbacks

of high-throughput localization experiments that are costly and time-consuming.

However, current subcellular localization predictors are limited in scope and

accuracy. In particular, most predictors perform well on certain locations or

with certain data sets while poorly on others. Here, we present PSI, a novel high

accuracy web server for plant subcellular localization prediction. PSI derives

the wisdom of multiple specialized predictors via a joint-approach of group

decision making strategy and machine learning methods to give an integrated best

result. The overall accuracy obtained (up to 93.4%) was higher than best

individual (CELLO) by ~10.7%. The precision of each predicable subcellular

location (more than 80%) far exceeds that of the individual predictors. It can

also deal with multi-localization proteins. PSI is expected to be a powerful tool

in protein location engineering as well as in plant sciences, while the strategy

employed could be applied to other integrative problems. A user-friendly web

server, PSI, has been developed for free access at http://bis.zju.edu.cn/psi/.

DOI: 10.1371/journal.pone.0075826

PMCID: PMC3806775

PMID: 24194827 [Indexed for MEDLINE]

1009. Bioinformatics. 2013 Oct 15;29(20):2651-2. doi: 10.1093/bioinformatics/btt457.

Epub 2013 Aug 13.

omiRas: a Web server for differential expression analysis of miRNAs derived from

small RNA-Seq data.

Müller S(1), Rycak L, Winter P, Kahl G, Koch I, Rotter B.

Author information:

(1)Plant Molecular Biology, Institute of Molecular BioSciences, University of

Frankfurt am Main, Marie-Curie-Str.9, 60439 Frankfurt am Main, GenXPro GmbH,

Frankfurt Innovation Center Biotechnology, Altenhöferallee 3, 60438 Frankfurt am

Main and Molecular Bioinformatics Group, Institute of Computer Science, Faculty

of Computer Science and Mathematics, Cluster of Excellence, Frankfurt

"Macromolecular Complexes", Robert-Mayer-Strasse 11-15, 60325 Frankfurt am Main,

Germany.

SUMMARY: Small RNA deep sequencing is widely used to characterize non-coding RNAs

(ncRNAs) differentially expressed between two conditions, e.g. healthy and

diseased individuals and to reveal insights into molecular mechanisms underlying

condition-specific phenotypic traits. The ncRNAome is composed of a multitude of

RNAs, such as transfer RNA, small nucleolar RNA and microRNA (miRNA), to name

few. Here we present omiRas, a Web server for the annotation, comparison and

visualization of interaction networks of ncRNAs derived from next-generation

sequencing experiments of two different conditions. The Web tool allows the user

to submit raw sequencing data and results are presented as: (i) static annotation

results including length distribution, mapping statistics, alignments and

quantification tables for each library as well as lists of differentially

expressed ncRNAs between conditions and (ii) an interactive network visualization

of user-selected miRNAs and their target genes based on the combination of

several miRNA-mRNA interaction databases.

AVAILABILITY AND IMPLEMENTATION: The omiRas Web server is implemented in Python,

PostgreSQL, R and can be accessed at: http://tools.genxpro.net/omiras/.

DOI: 10.1093/bioinformatics/btt457

PMID: 23946503 [Indexed for MEDLINE]

1010. Bioinformatics. 2013 Oct 15;29(20):2649-50. doi: 10.1093/bioinformatics/btt441.

Epub 2013 Aug 7.

WebRASP: a server for computing energy scores to assess the accuracy and

stability of RNA 3D structures.

Norambuena T(1), Cares JF, Capriotti E, Melo F.

Author information:

(1)Facultad de Ciencias Biologicas, Departamento de Genetica Molecular y

Microbiologia, Pontificia Universidad Catolica de Chile, Alameda 340, Molecular

Bioinformatics Laboratory, Millennium Institute on Immunology and Immunotherapy,

Santiago, Chile and Division of Informatics, Department of Pathology, University

of Alabama at Birmingham, 619 19 st. south, Birmingham, AL 35249, USA.

SUMMARY: The understanding of the biological role of RNA molecules has changed.

Although it is widely accepted that RNAs play important regulatory roles without

necessarily coding for proteins, the functions of many of these non-coding RNAs

are unknown. Thus, determining or modeling the 3D structure of RNA molecules as

well as assessing their accuracy and stability has become of great importance for

characterizing their functional activity. Here, we introduce a new web

application, WebRASP, that uses knowledge-based potentials for scoring RNA

structures based on distance-dependent pairwise atomic interactions. This web

server allows the users to upload a structure in PDB format, select several

options to visualize the structure and calculate the energy profile. The server

contains online help, tutorials and links to other related resources. We believe

this server will be a useful tool for predicting and assessing the quality of RNA

3D structures.

AVAILABILITY AND IMPLEMENTATION: The web server is available at

http://melolab.org/webrasp. It has been tested on the most popular web browsers

and requires Java plugin for Jmol visualization.

DOI: 10.1093/bioinformatics/btt441

PMCID: PMC3789544

PMID: 23929030 [Indexed for MEDLINE]

1011. Bioinformatics. 2013 Oct 15;29(20):2647-8. doi: 10.1093/bioinformatics/btt451.

Epub 2013 Aug 5.

Pclust: protein network visualization highlighting experimental data.

Li W(1), Kinch LN, Grishin NV.

Author information:

(1)Departments of Biophysics and Biochemistry and Howard Hughes Medical

Institute, University of Texas Southwestern Medical Center, Dallas, TX

75390-9050, USA.

SUMMARY: One approach to infer functions of new proteins from their homologs

utilizes visualization of an all-against-all pairwise similarity network (A2ApsN)

that exploits the speed of BLAST and avoids the complexity of multiple sequence

alignment. However, identifying functions of the protein clusters in A2ApsN is

never trivial, due to a lack of linking characterized proteins to their relevant

information in current software packages. Given the database errors introduced by

automatic annotation transfer, functional deduction should be made from proteins

with experimental studies, i.e. 'reference proteins'. Here, we present a web

server, termed Pclust, which provides a user-friendly interface to visualize the

A2ApsN, placing emphasis on such 'reference proteins' and providing access to

their full information in source databases, e.g. articles in PubMed. The

identification of 'reference proteins' and the ease of cross-database linkage

will facilitate understanding the functions of protein clusters in the network,

thus promoting interpretation of proteins of interest.

AVAILABILITY: The Pclust server is freely available at

http://prodata.swmed.edu/pclust

DOI: 10.1093/bioinformatics/btt451

PMCID: PMC3789550

PMID: 23918248 [Indexed for MEDLINE]

1012. Circ Res. 2013 Oct 12;113(9):1043-53. doi: 10.1161/CIRCRESAHA.113.301151. Epub

2013 Aug 21.

Integration of cardiac proteome biology and medicine by a specialized

knowledgebase.

Zong NC(1), Li H, Li H, Lam MP, Jimenez RC, Kim CS, Deng N, Kim AK, Choi JH,

Zelaya I, Liem D, Meyer D, Odeberg J, Fang C, Lu HJ, Xu T, Weiss J, Duan H, Uhlen

M, Yates JR 3rd, Apweiler R, Ge J, Hermjakob H, Ping P.

Author information:

(1)From the NHLBI Proteomics Center at UCLA/NHLBI Proteomics Program.

RATIONALE: Omics sciences enable a systems-level perspective in characterizing

cardiovascular biology. Integration of diverse proteomics data via a

computational strategy will catalyze the assembly of contextualized knowledge,

foster discoveries through multidisciplinary investigations, and minimize

unnecessary redundancy in research efforts.

OBJECTIVE: The goal of this project is to develop a consolidated cardiac proteome

knowledgebase with novel bioinformatics pipeline and Web portals, thereby serving

as a new resource to advance cardiovascular biology and medicine.

METHODS AND RESULTS: We created Cardiac Organellar Protein Atlas Knowledgebase

(COPaKB; www.HeartProteome.org), a centralized platform of high-quality cardiac

proteomic data, bioinformatics tools, and relevant cardiovascular phenotypes.

Currently, COPaKB features 8 organellar modules, comprising 4203 LC-MS/MS

experiments from human, mouse, drosophila, and Caenorhabditis elegans, as well as

expression images of 10,924 proteins in human myocardium. In addition, the

Java-coded bioinformatics tools provided by COPaKB enable cardiovascular

investigators in all disciplines to retrieve and analyze pertinent organellar

protein properties of interest.

CONCLUSIONS: COPaKB provides an innovative and interactive resource that connects

research interests with the new biological discoveries in protein sciences. With

an array of intuitive tools in this unified Web server, nonproteomics

investigators can conveniently collaborate with proteomics specialists to dissect

the molecular signatures of cardiovascular phenotypes.

DOI: 10.1161/CIRCRESAHA.113.301151

PMCID: PMC4076475

PMID: 23965338 [Indexed for MEDLINE]

1013. PLoS One. 2013 Oct 9;8(10):e75726. doi: 10.1371/journal.pone.0075726. eCollection

2013.

AcalPred: a sequence-based tool for discriminating between acidic and alkaline

enzymes.

Lin H(1), Chen W, Ding H.

Author information:

(1)Key Laboratory for NeuroInformation of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu, China.

The structure and activity of enzymes are influenced by pH value of their

surroundings. Although many enzymes work well in the pH range from 6 to 8, some

specific enzymes have good efficiencies only in acidic (pH<5) or alkaline (pH>9)

solution. Studies have demonstrated that the activities of enzymes correlate with

their primary sequences. It is crucial to judge enzyme adaptation to acidic or

alkaline environment from its amino acid sequence in molecular mechanism

clarification and the design of high efficient enzymes. In this study, we

developed a sequence-based method to discriminate acidic enzymes from alkaline

enzymes. The analysis of variance was used to choose the optimized discriminating

features derived from g-gap dipeptide compositions. And support vector machine

was utilized to establish the prediction model. In the rigorous jackknife

cross-validation, the overall accuracy of 96.7% was achieved. The method can

correctly predict 96.3% acidic and 97.1% alkaline enzymes. Through the comparison

between the proposed method and previous methods, it is demonstrated that the

proposed method is more accurate. On the basis of this proposed method, we have

built an online web-server called AcalPred which can be freely accessed from the

website (http://lin.uestc.edu.cn/server/AcalPred). We believe that the AcalPred

will become a powerful tool to study enzyme adaptation to acidic or alkaline

environment.

DOI: 10.1371/journal.pone.0075726

PMCID: PMC3794003

PMID: 24130738 [Indexed for MEDLINE]

1014. PeerJ. 2013 Oct 3;1:e171. doi: 10.7717/peerj.171. eCollection 2013.

iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for

predicting cysteine S-nitrosylation sites in proteins.

Xu Y(1), Shao XJ, Wu LY, Deng NY, Chou KC.

Author information:

(1)Department of Information and Computer Science, University of Science and

Technology Beijing , Beijing , China ; Gordon Life Science Institute , Belmont,

MA , USA.

As one of the most important and universal posttranslational modifications (PTMs)

of proteins, S-nitrosylation (SNO) plays crucial roles in a variety of biological

processes, including the regulation of cellular dynamics and many signaling

events. Knowledge of SNO sites in proteins is very useful for drug development

and basic research as well. Unfortunately, it is both time-consuming and costly

to determine the SNO sites purely based on biological experiments. Facing the

explosive protein sequence data generated in the post-genomic era, we are

challenged to develop automated vehicles for timely and effectively determining

the SNO sites for uncharacterized proteins. To address the challenge, a new

predictor called iSNO-AAPair was developed by taking into account the coupling

effects for all the pairs formed by the nearest residues and the pairs by the

next nearest residues along protein chains. The cross-validation results on a

state-of-the-art benchmark have shown that the new predictor outperformed the

existing predictors. The same was true when tested by the independent proteins

whose experimental SNO sites were known. A user-friendly web-server for

iSNO-AAPair was established at http://app.aporc.org/iSNO-AAPair/, by which users

can easily obtain their desired results without the need to follow the

mathematical equations involved during its development.

DOI: 10.7717/peerj.171

PMCID: PMC3792191

PMID: 24109555

1015. Eur J Hum Genet. 2013 Oct;21(10):1128-33. doi: 10.1038/ejhg.2013.7. Epub 2013 Mar

6.

A global map for dissecting phenotypic variants in human lincRNAs.

Ning S(1), Wang P, Ye J, Li X, Li R, Zhao Z, Huo X, Wang L, Li F, Li X.

Author information:

(1)College of Bioinformatics Science and Technology, Harbin Medical University,

Harbin, China.

Large intergenic noncoding RNAs (lincRNAs) are emerging as key factors of

multiple cellular processes. Cumulative evidence has linked lincRNA polymorphisms

to diverse diseases. However, the global properties of lincRNA polymorphisms and

their implications for human disease remain largely unknown. Here we performed a

systematic analysis of naturally occurring variants in human lincRNAs, with a

particular focus on lincRNA polymorphism as novel risk factor of disease

etiology. We found that lincRNAs exhibited a relatively low level of

polymorphisms, and low single-nucleotide polymorphism (SNP) density lincRNAs

might have a broad range of functions. We also found that some polymorphisms in

evolutionarily conserved regions of lincRNAs had significant effects on predicted

RNA secondary structures, indicating their potential contribution to diseases. We

mapped currently available phenotype-associated SNPs to lincRNAs and found that

lincRNAs were associated with a wide range of human diseases. Some lincRNAs could

be responsible for particular diseases. Our results provided not only a global

perspective on genetic variants in human lincRNAs but also novel insights into

the function and etiology of lincRNA. All the data in this study can be accessed

and retrieved freely via a web server at http://bioinfo.hrbmu.edu.cn/lincPoly.

DOI: 10.1038/ejhg.2013.7

PMCID: PMC3778363

PMID: 23463026 [Indexed for MEDLINE]

1016. Protein Eng Des Sel. 2013 Oct;26(10):631-4. doi: 10.1093/protein/gzt018. Epub

2013 May 9.

EpiDOCK: a molecular docking-based tool for MHC class II binding prediction.

Atanasova M(1), Patronov A, Dimitrov I, Flower DR, Doytchinova I.

Author information:

(1)School of Pharmacy, Medical University of Sofia, 2 Dunav street, 1000 Sofia,

Bulgaria.

Cellular peptide vaccines contain T-cell epitopes. The main prerequisite for a

peptide to act as a T-cell epitope is that it binds to a major histocompatibility

complex (MHC) protein. Peptide MHC binder identification is an extremely costly

experimental challenge since human MHCs, named human leukocyte antigen, are

highly polymorphic and polygenic. Here we present EpiDOCK, the first

structure-based server for MHC class II binding prediction. EpiDOCK predicts

binding to the 23 most frequent human, MHC class II proteins. It identifies 90%

of true binders and 76% of true non-binders, with an overall accuracy of 83%.

EpiDOCK is freely accessible at http://epidock.ddg-pharmfac.net.

DOI: 10.1093/protein/gzt018

PMID: 23661105 [Indexed for MEDLINE]

1017. Proteins. 2013 Oct;81(10):1823-39. doi: 10.1002/prot.24327. Epub 2013 Aug 19.

Adaptive Smith-Waterman residue match seeding for protein structural alignment.

Topham CM(1), Rouquier M, Tarrat N, André I.

Author information:

(1)Université de Toulouse, INSA, UPS, INP, LISBP, 135 Avenue de Rangueil,

F-31077, Toulouse, France; CNRS, UMR5504, F-31400, Toulouse, France; INRA, UMR792

Ingénierie des Systèmes Biologiques et des Procédés, F-31400, Toulouse, France.

The POLYFIT rigid-body algorithm for automated global pairwise and multiple

protein structural alignment is presented. Smith-Waterman local alignment is used

to establish a set of seed equivalences that are extended using Needleman-Wunsch

dynamic programming techniques. Structural and functional interaction constraints

provided by evolution are encoded as one-dimensional residue physical environment

strings for alignment of highly structurally overlapped protein pairs. Local

structure alignment of more distantly related pairs is carried out using

rigid-body conformational matching of 15-residue fragments, with allowance made

for less stringent conformational matching of metal-ion and small molecule

ligand-contact, disulphide bridge, and cis-peptide correspondences. Protein

structural plasticity is accommodated through the stepped adjustment of a single

empirical distance parameter value in the calculation of the Smith-Waterman

dynamic programming matrix. Structural overlap is used both as a measure of

similarity and to assess alignment quality. Pairwise alignment accuracy has been

benchmarked against that of 10 widely used aligners on the Sippl and Wiederstein

set of difficult pairwise structure alignment problems, and more extensively

against that of Matt, SALIGN, and MUSTANG in pairwise and multiple structural

alignments of protein domains with low shared sequence identity in the

SCOP-ASTRAL 40% compendium. The results demonstrate the advantages of POLYFIT

over other aligners in the efficient and robust identification of matching seed

residue positions in distantly related protein targets and in the generation of

longer structurally overlapped alignment lengths. Superposition-based application

areas include comparative modeling and protein and ligand design. POLYFIT is

available on the Web server at http://polyfit.insa-toulouse.fr.

Copyright © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24327

PMID: 23720362 [Indexed for MEDLINE]

1018. BMC Genomics. 2013 Sep 30;14:664. doi: 10.1186/1471-2164-14-664.

Bolbase: a comprehensive genomics database for Brassica oleracea.

Yu J(1), Zhao M, Wang X, Tong C, Huang S, Tehrim S, Liu Y, Hua W, Liu S.

Author information:

(1)The Key Laboratory of Oil Crops Biology and Genetic Breeding, the Ministry of

Agriculture, Oil Crops Research Institute, the Chinese Academy of Agricultural

Sciences, Wuhan 430062, China. huawei@oilcrops.cn.

BACKGROUND: Brassica oleracea is a morphologically diverse species in the family

Brassicaceae and contains a group of nutrition-rich vegetable crops, including

common heading cabbage, cauliflower, broccoli, kohlrabi, kale, Brussels sprouts.

This diversity along with its phylogenetic membership in a group of three diploid

and three tetraploid species, and the recent availability of genome sequences

within Brassica provide an unprecedented opportunity to study intra- and

inter-species divergence and evolution in this species and its close relatives.

DESCRIPTION: We have developed a comprehensive database, Bolbase, which provides

access to the B. oleracea genome data and comparative genomics information. The

whole genome of B. oleracea is available, including nine fully assembled

chromosomes and 1,848 scaffolds, with 45,758 predicted genes, 13,382 transposable

elements, and 3,581 non-coding RNAs. Comparative genomics information is

available, including syntenic regions among B. oleracea, Brassica rapa and

Arabidopsis thaliana, synonymous (Ks) and non-synonymous (Ka) substitution rates

between orthologous gene pairs, gene families or clusters, and differences in

quantity, category, and distribution of transposable elements on chromosomes.

Bolbase provides useful search and data mining tools, including a keyword search,

a local BLAST server, and a customized GBrowse tool, which can be used to extract

annotations of genome components, identify similar sequences and visualize

syntenic regions among species. Users can download all genomic data and explore

comparative genomics in a highly visual setting.

CONCLUSIONS: Bolbase is the first resource platform for the B. oleracea genome

and for genomic comparisons with its relatives, and thus it will help the

research community to better study the function and evolution of Brassica genomes

as well as enhance molecular breeding research. This database will be updated

regularly with new features, improvements to genome annotation, and new genomic

sequences as they become available. Bolbase is freely available at

http://ocri-genomics.org/bolbase.

DOI: 10.1186/1471-2164-14-664

PMCID: PMC3849793

PMID: 24079801 [Indexed for MEDLINE]

1019. BMC Bioinformatics. 2013 Sep 25;14:284. doi: 10.1186/1471-2105-14-284.

DaGO-Fun: tool for Gene Ontology-based functional analysis using term information

content measures.

Mazandu GK(1), Mulder NJ.

Author information:

(1)Computational Biology Group, Department of Clinical Laboratory Sciences,

Institute of Infectious Disease and Molecular Medicine, University of Cape Town,

Medical School, Observatory, Cape Town, 7925, South Africa.

gmazandu@cbio.uct.ac.za.

BACKGROUND: The use of Gene Ontology (GO) data in protein analyses have largely

contributed to the improved outcomes of these analyses. Several GO semantic

similarity measures have been proposed in recent years and provide tools that

allow the integration of biological knowledge embedded in the GO structure into

different biological analyses. There is a need for a unified tool that provides

the scientific community with the opportunity to explore these different GO

similarity measure approaches and their biological applications.

RESULTS: We have developed DaGO-Fun, an online tool available at

http://web.cbio.uct.ac.za/ITGOM, which incorporates many different GO similarity

measures for exploring, analyzing and comparing GO terms and proteins within the

context of GO. It uses GO data and UniProt proteins with their GO annotations as

provided by the Gene Ontology Annotation (GOA) project to precompute GO term

information content (IC), enabling rapid response to user queries.

CONCLUSIONS: The DaGO-Fun online tool presents the advantage of integrating all

the relevant IC-based GO similarity measures, including topology- and

annotation-based approaches to facilitate effective exploration of these

measures, thus enabling users to choose the most relevant approach for their

application. Furthermore, this tool includes several biological applications

related to GO semantic similarity scores, including the retrieval of genes based

on their GO annotations, the clustering of functionally related genes within a

set, and term enrichment analysis.

DOI: 10.1186/1471-2105-14-284

PMCID: PMC3849277

PMID: 24067102 [Indexed for MEDLINE]

1020. Bioinformatics. 2013 Sep 15;29(18):2360-2. doi: 10.1093/bioinformatics/btt401.

Epub 2013 Jul 9.

iLoops: a protein-protein interaction prediction server based on structural

features.

Planas-Iglesias J(1), Marin-Lopez MA, Bonet J, Garcia-Garcia J, Oliva B.

Author information:

(1)Structural Bioinformatics Laboratory, Universitat Pompeu Fabra, 08003

Barcelona, Spain.

SUMMARY: Protein-protein interactions play a critical role in many biological

processes. Despite that, the number of servers that provide an easy and

comprehensive method to predict them is still limited. Here, we present iLoops, a

web server that predicts whether a pair of proteins can interact using local

structural features. The inputs of the server are as follows: (i) the sequences

of the query proteins and (ii) the pairs to be tested. Structural features are

assigned to the query proteins by sequence similarity. Pairs of structural

features (formed by loops or domains) are classified according to their

likelihood to favor or disfavor a protein-protein interaction, depending on their

observation in known interacting and non-interacting pairs. The server evaluates

the putative interaction using a random forest classifier.

AVAILABILITY: iLoops is available at http://sbi.imim.es/iLoops.php

CONTACT: baldo.oliva@upf.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt401

PMID: 23842807 [Indexed for MEDLINE]

1021. Bioinformatics. 2013 Sep 15;29(18):2357-9. doi: 10.1093/bioinformatics/btt399.

Epub 2013 Jul 9.

Allosite: a method for predicting allosteric sites.

Huang W(1), Lu S, Huang Z, Liu X, Mou L, Luo Y, Zhao Y, Liu Y, Chen Z, Hou T,

Zhang J.

Author information:

(1)Department of Pathophysiology, Shanghai Jiao-Tong University School of

Medicine, Shanghai 200025, China.

MOTIVATION: The use of allosteric modulators as preferred therapeutic agents

against classic orthosteric ligands has colossal advantages, including higher

specificity, fewer side effects and lower toxicity. Therefore, the computational

prediction of allosteric sites in proteins is receiving increased attention in

the field of drug discovery. Allosite is a newly developed automatic tool for the

prediction of allosteric sites in proteins of interest and is now available

through a web server.

AVAILABILITY: The Allosite server and tutorials are freely available at

http://mdl.shsmu.edu.cn/AST CONTACT: jian.zhang@sjtu.edu.cn

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt399

PMID: 23842804 [Indexed for MEDLINE]

1022. Bioinformatics. 2013 Sep 15;29(18):2285-91. doi: 10.1093/bioinformatics/btt369.

Epub 2013 Jun 26.

Prediction of site-specific interactions in antibody-antigen complexes: the

proABC method and server.

Olimpieri PP(1), Chailyan A, Tramontano A, Marcatili P.

Author information:

(1)Department of Physics, Sapienza University and Istituto Pasteur - Fondazione

Cenci Bolognetti, 00185 Rome, Italy.

MOTIVATION: Antibodies or immunoglobulins are proteins of paramount importance in

the immune system. They are extremely relevant as diagnostic, biotechnological

and therapeutic tools. Their modular structure makes it easy to re-engineer them

for specific purposes. Short of undergoing a trial and error process, these

experiments, as well as others, need to rely on an understanding of the specific

determinants of the antibody binding mode.

RESULTS: In this article, we present a method to identify, on the basis of the

antibody sequence alone, which residues of an antibody directly interact with its

cognate antigen. The method, based on the random forest automatic learning

techniques, reaches a recall and specificity as high as 80% and is implemented as

a free and easy-to-use server, named prediction of Antibody Contacts. We believe

that it can be of great help in re-design experiments as well as a guide for

molecular docking experiments. The results that we obtained also allowed us to

dissect which features of the antibody sequence contribute most to the

involvement of specific residues in binding to the antigen.

AVAILABILITY: http://www.biocomputing.it/proABC.

CONTACT: anna.tramontano@uniroma1.it or paolo.marcatili@gmail.com

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt369

PMCID: PMC3753563

PMID: 23803466 [Indexed for MEDLINE]

1023. PLoS One. 2013 Sep 13;8(9):e73957. doi: 10.1371/journal.pone.0073957. eCollection

2013.

In silico approach for predicting toxicity of peptides and proteins.

Gupta S(1), Kapoor P, Chaudhary K, Gautam A, Kumar R; Open Source Drug Discovery

Consortium, Raghava GP.

Author information:

(1)Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh,

India.

BACKGROUND: Over the past few decades, scientific research has been focused on

developing peptide/protein-based therapies to treat various diseases. With the

several advantages over small molecules, including high specificity, high

penetration, ease of manufacturing, peptides have emerged as promising

therapeutic molecules against many diseases. However, one of the bottlenecks in

peptide/protein-based therapy is their toxicity. Therefore, in the present study,

we developed in silico models for predicting toxicity of peptides and proteins.

DESCRIPTION: We obtained toxic peptides having 35 or fewer residues from various

databases for developing prediction models. Non-toxic or random peptides were

obtained from SwissProt and TrEMBL. It was observed that certain residues like

Cys, His, Asn, and Pro are abundant as well as preferred at various positions in

toxic peptides. We developed models based on machine learning technique and

quantitative matrix using various properties of peptides for predicting toxicity

of peptides. The performance of dipeptide-based model in terms of accuracy was

94.50% with MCC 0.88. In addition, various motifs were extracted from the toxic

peptides and this information was combined with dipeptide-based model for

developing a hybrid model. In order to evaluate the over-optimization of the best

model based on dipeptide composition, we evaluated its performance on independent

datasets and achieved accuracy around 90%. Based on above study, a web server,

ToxinPred has been developed, which would be helpful in predicting (i) toxicity

or non-toxicity of peptides, (ii) minimum mutations in peptides for increasing or

decreasing their toxicity, and (iii) toxic regions in proteins.

CONCLUSION: ToxinPred is a unique in silico method of its kind, which will be

useful in predicting toxicity of peptides/proteins. In addition, it will be

useful in designing least toxic peptides and discovering toxic regions in

proteins. We hope that the development of ToxinPred will provide momentum to

peptide/protein-based drug discovery (http://crdd.osdd.net/raghava/toxinpred/).

DOI: 10.1371/journal.pone.0073957

PMCID: PMC3772798

PMID: 24058508 [Indexed for MEDLINE]

1024. PLoS One. 2013 Sep 6;8(9):e74092. doi: 10.1371/journal.pone.0074092. eCollection

2013.

GOMoDo: A GPCRs online modeling and docking webserver.

Sandal M(1), Duy TP, Cona M, Zung H, Carloni P, Musiani F, Giorgetti A.

Author information:

(1)Department of Biotechnology, University of Verona, Ca' Vignal 1, Verona,

Italy.

G-protein coupled receptors (GPCRs) are a superfamily of cell signaling membrane

proteins that include >750 members in the human genome alone. They are the

largest family of drug targets. The vast diversity and relevance of GPCRs

contrasts with the paucity of structures available: only 21 unique GPCR

structures have been experimentally determined as of the beginning of 2013.

User-friendly modeling and small molecule docking tools are thus in great demand.

While both GPCR structural predictions and docking servers exist separately, with

GOMoDo (GPCR Online Modeling and Docking), we provide a web server to seamlessly

model GPCR structures and dock ligands to the models in a single consistent

pipeline. GOMoDo can automatically perform template choice, homology modeling and

either blind or information-driven docking by combining together proven, state of

the art bioinformatic tools. The web server gives the user the possibility of

guiding the whole procedure. The GOMoDo server is freely accessible at

http://molsim.sci.univr.it/gomodo.

DOI: 10.1371/journal.pone.0074092

PMCID: PMC3772745

PMID: 24058518 [Indexed for MEDLINE]

1025. PLoS One. 2013 Sep 3;8(9):e74002. doi: 10.1371/journal.pone.0074002. eCollection

2013.

Systematic analysis and prediction of pupylation sites in prokaryotic proteins.

Chen X(1), Qiu JD, Shi SP, Suo SB, Liang RP.

Author information:

(1)Department of Chemistry, Nanchang University, Nanchang, P.R. China.

Prokaryotic ubiquitin-like protein (Pup) is the first identified prokaryotic

protein that is functionally analogous to ubiquitin. Recent studies have shed

light on the Pup activation and conjugation to target proteins to be a signal for

the selective degradation proteins in Mycobacterium tuberculosis (Mtb). By

covalently conjugating the Pup, pupylation functions as a critical

post-translational modification (PTM) conserved in actinomycetes. Detecting

pupylation sites is crucial and fundamental for understanding the molecular

mechanisms of Pup. Yet comparative studies with other PTM suggest that the

development of accurate and complete repertories of pupylation is still in its

early stages. Unbiased screening for pupylation sites by experimental methods is

time consuming and expensive; in silico prediction can provide highly potential

candidates and reduce the number of potential candidates that require further in

vivo or in vitro confirmation. Here, we present an effective classifier of

PupPred for predicting pupylation sites, which shows better performance than

existing classifiers. Importantly, this work not only investigates the

sequential, structural and evolutionary hallmarks around pupylation sites but

also compares the differences of pupylation and ubiquitylation from the

environmental, conservative and functional characterization of substrates. These

prediction and analysis results may be helpful for further experimental

investigation of degradation proteins in prokaryotes. Finally, the PupPred server

is available at http://bioinfo.ncu.edu.cn/PupPred.aspx.

DOI: 10.1371/journal.pone.0074002

PMCID: PMC3760804

PMID: 24019945 [Indexed for MEDLINE]

1026. Bioinformatics. 2013 Sep 1;29(17):2197-8. doi: 10.1093/bioinformatics/btt356.

Epub 2013 Jun 19.

ChroMoS: an integrated web tool for SNP classification, prioritization and

functional interpretation.

Barenboim M(1), Manke T.

Author information:

(1)Bioinformatics and Deep-Sequencing Unit, Max-Planck Institute of Immunobiology

and Epigenetics, Freiburg, Germany. barenboim@ie-freiburg.mpg.de

Genome-wide association studies and re-sequencing projects are revealing an

increasing number of disease-associated SNPs, a large fraction of which are

non-coding. Although they could have relevance for disease susceptibility and

progression, the lack of information about regulatory regions impedes the

assessment of their functionality. Here we present a web server, ChroMoS

(Chromatin Modified SNPs), which combines genetic and epigenetic data with the

goal of facilitating SNPs' classification, prioritization and prediction of their

functional consequences. ChroMoS uses a large database of SNPs and chromatin

states, but allows a user to provide his/her own genetic information. Based on

the SNP classification and interactive prioritization, a user can compute the

functional impact of multiple SNPs using two prediction tools, one for

differential analysis of transcription factor binding (sTRAP) and another for

SNPs with potential impact on binding of miRNAs (MicroSNiPer).AVAILABILITY: Web

server, ChroMoS, is freely available at

http://epicenter.immunbio.mpg.de/services/chromos.

DOI: 10.1093/bioinformatics/btt356

PMCID: PMC3740627

PMID: 23782616 [Indexed for MEDLINE]

1027. Genetics. 2013 Sep;195(1):37-45. doi: 10.1534/genetics.113.151340. Epub 2013 Jun

21.

UP-TORR: online tool for accurate and Up-to-Date annotation of RNAi Reagents.

Hu Y(1), Roesel C, Flockhart I, Perkins L, Perrimon N, Mohr SE.

Author information:

(1)Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115.

Comment in

Genetics. 2013 Sep;195(1):7-8.

G3 (Bethesda). 2013 Sep;3(9):1451-2.

RNA interference (RNAi) is a widely adopted tool for loss-of-function studies but

RNAi results only have biological relevance if the reagents are appropriately

mapped to genes. Several groups have designed and generated RNAi reagent

libraries for studies in cells or in vivo for Drosophila and other species. At

first glance, matching RNAi reagents to genes appears to be a simple problem, as

each reagent is typically designed to target a single gene. In practice, however,

the reagent-gene relationship is complex. Although the sequences of

oligonucleotides used to generate most types of RNAi reagents are static, the

reference genome and gene annotations are regularly updated. Thus, at the time a

researcher chooses an RNAi reagent or analyzes RNAi data, the most current

interpretation of the RNAi reagent-gene relationship, as well as related

information regarding specificity (e.g., predicted off-target effects), can be

different from the original interpretation. Here, we describe a set of strategies

and an accompanying online tool, UP-TORR (for Updated Targets of RNAi Reagents;

www.flyrnai.org/up-torr), useful for accurate and up-to-date annotation of

cell-based and in vivo RNAi reagents. Importantly, UP-TORR automatically

synchronizes with gene annotations daily, retrieving the most current information

available, and for Drosophila, also synchronizes with the major reagent

collections. Thus, UP-TORR allows users to choose the most appropriate RNAi

reagents at the onset of a study, as well as to perform the most appropriate

analyses of results of RNAi-based studies.

DOI: 10.1534/genetics.113.151340

PMCID: PMC3761311

PMID: 23792952 [Indexed for MEDLINE]

1028. J Environ Manage. 2013 Sep;127 Suppl:S168-83. doi: 10.1016/j.jenvman.2013.02.051.

Epub 2013 Apr 10.

LandCaRe DSS--an interactive decision support system for climate change impact

assessment and the analysis of potential agricultural land use adaptation

strategies.

Wenkel KO(1), Berg M, Mirschel W, Wieland R, Nendel C, Köstner B.

Author information:

(1)Leibniz-Centre for Agricultural Landscape Research, Institute of Landscape

Systems Analysis, Eberswalder Straße 84, Müncheberg, Germany. wenkel@zalf.de

Decision support to develop viable climate change adaptation strategies for

agriculture and regional land use management encompasses a wide range of options

and issues. Up to now, only a few suitable tools and methods have existed for

farmers and regional stakeholders that support the process of decision-making in

this field. The interactive model-based spatial information and decision support

system LandCaRe DSS attempts to close the existing methodical gap. This system

supports interactive spatial scenario simulations, multi-ensemble and multi-model

simulations at the regional scale, as well as the complex impact assessment of

potential land use adaptation strategies at the local scale. The system is

connected to a local geo-database and via the internet to a climate data server.

LandCaRe DSS uses a multitude of scale-specific ecological impact models, which

are linked in various ways. At the local scale (farm scale), biophysical models

are directly coupled with a farm economy calculator. New or alternative

simulation models can easily be added, thanks to the innovative architecture and

design of the DSS. Scenario simulations can be conducted with a reasonable amount

of effort. The interactive LandCaRe DSS prototype also offers a variety of data

analysis and visualisation tools, a help system for users and a farmer

information system for climate adaptation in agriculture. This paper presents the

theoretical background, the conceptual framework, and the structure and

methodology behind LandCaRe DSS. Scenario studies at the regional and local scale

for the two Eastern German regions of Uckermark (dry lowlands, 2600 km(2)) and

Weißeritz (humid mountain area, 400 km(2)) were conducted in close cooperation

with stakeholders to test the functionality of the DSS prototype. The system is

gradually being transformed into a web version (http://www.landcare-dss.de) to

ensure the broadest possible distribution of LandCaRe DSS to the public. The

system will be continuously developed, updated and used in different research

projects and as a learning and knowledge-sharing tool for students. The main

objective of LandCaRe DSS is to provide information on the complex long-term

impacts of climate change and on potential management options for adaptation by

answering "what-if" type questions.

Copyright © 2013 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jenvman.2013.02.051

PMID: 23582740 [Indexed for MEDLINE]

1029. J Ultrasound Med. 2013 Sep;32(9):1601-5. doi: 10.7863/ultra.32.9.1601.

Formative assessment based on an audit and feedback improves nuchal translucency

ultrasound image quality.

Chalouhi GE(1), Salomon LJ, Fontanges M, Althuser M, Haddad G, Scemama O, Chabot

JM, Duyme M, Fries N.

Author information:

(1)Department of Obstetrics and Fetal Medicine, Paris Descartes University,

Assistance Publique-Hôpitaux de Paris, Hôpital Necker-Enfants-Malades, Paris,

France.

OBJECTIVES: The purpose of this work was to study the impact of an audit and

feedback on the quality of routine first-trimester nuchal transparency ultrasound

images.

METHODS: Eighty-eight sonographers were each sent 2 different series of 30

consecutive nuchal translucency images at a mean interval of 3 months to a

dedicated, protected server for remote double-blind independent analysis based on

the new Collège Français d'Echographie Foetale/Centre National de la Recherche

Scientifique image-scoring method

(https://www.cfef.org/evaluation/ISMCFEFCNRS.pdf). The sonographers were

classified as low (score below the median) or high (score above the median)

scorers for each series. Before their second evaluation, 73 of the 88

sonographers received a feedback report on their first series of images, whereas

the other 15 participants received no feedback. The baseline characteristics of

the participants who did and did not receive feedback were comparable.

RESULTS: Participants who received feedback increased their average score

significantly, from a mean ± SD of 11.1 ± 1.3 to 13.4 ± 1.4 among low scorers (P

< .00001) and from 15.1 ± 1.2 to 16.0 ± 1.4 among high scorers (P < .001),

whereas no significant change was seen among participants who received no

feedback (low scorers, 10.9 ± 1.5 to 12.1 ± 2.0; P = .11; high scorers, 14.7 ±

1.3 to 14.6 ± 1.3; P = .99). The proportion of satisfactory images increased by

48% among low scorers who received feedback.

CONCLUSIONS: Formative assessment based on a moderately intensive audit and

feedback is feasible and effective for improving the quality of routine

first-trimester nuchal transparency ultrasound images.

DOI: 10.7863/ultra.32.9.1601

PMID: 23980221 [Indexed for MEDLINE]

1030. Nucleic Acids Res. 2013 Sep;41(17):8034-44. doi: 10.1093/nar/gkt606. Epub 2013

Jul 17.

CRISPRmap: an automated classification of repeat conservation in prokaryotic

adaptive immune systems.

Lange SJ(1), Alkhnbashi OS, Rose D, Will S, Backofen R.

Author information:

(1)Bioinformatics Group, Department of Computer Science,

Albert-Ludwigs-University Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg,

Germany, ZBSA Centre for Biological Systems Analysis, Albert-Ludwigs-University

Freiburg, Habsburgerstr. 49, 79104 Freiburg, Germany, BIOSS Centre for Biological

Signalling Studies, Cluster of Excellence, Albert-Ludwigs-University Freiburg,

Germany and Center for non-coding RNA in Technology and Health, University of

Copenhagen, Gronnegardsvej 3, DK-1870 Frederiksberg C, Denmark.

Central to Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-Cas

systems are repeated RNA sequences that serve as Cas-protein-binding templates.

Classification is based on the architectural composition of associated Cas

proteins, considering repeat evolution is essential to complete the picture. We

compiled the largest data set of CRISPRs to date, performed comprehensive,

independent clustering analyses and identified a novel set of 40 conserved

sequence families and 33 potential structure motifs for Cas-endoribonucleases

with some distinct conservation patterns. Evolutionary relationships are

presented as a hierarchical map of sequence and structure similarities for both a

quick and detailed insight into the diversity of CRISPR-Cas systems. In a

comparison with Cas-subtypes, I-C, I-E, I-F and type II were strongly coupled and

the remaining type I and type III subtypes were loosely coupled to repeat and

Cas1 evolution, respectively. Subtypes with a strong link to CRISPR evolution

were almost exclusive to bacteria; nevertheless, we identified rare examples of

potential horizontal transfer of I-C and I-E systems into archaeal organisms. Our

easy-to-use web server provides an automated assignment of newly sequenced

CRISPRs to our classification system and enables more informed choices on future

hypotheses in CRISPR-Cas research:

http://rna.informatik.uni-freiburg.de/CRISPRmap.

DOI: 10.1093/nar/gkt606

PMCID: PMC3783184

PMID: 23863837 [Indexed for MEDLINE]

1031. Nucleic Acids Res. 2013 Sep;41(16):7606-14. doi: 10.1093/nar/gkt544. Epub 2013

Jun 20.

Novel approach for selecting the best predictor for identifying the binding sites

in DNA binding proteins.

Nagarajan R(1), Ahmad S, Gromiha MM.

Author information:

(1)Department of Biotechnology, Indian Institute of Technology Madras, Chennai

600036, India and National Institute of Biomedical Innovation, Osaka, Japan.

Protein-DNA complexes play vital roles in many cellular processes by the

interactions of amino acids with DNA. Several computational methods have been

developed for predicting the interacting residues in DNA-binding proteins using

sequence and/or structural information. These methods showed different levels of

accuracies, which may depend on the choice of data sets used in training, the

feature sets selected for developing a predictive model, the ability of the

models to capture information useful for prediction or a combination of these

factors. In many cases, different methods are likely to produce similar results,

whereas in others, the predictors may return contradictory predictions. In this

situation, a priori estimates of prediction performance applicable to the system

being investigated would be helpful for biologists to choose the best method for

designing their experiments. In this work, we have constructed unbiased,

stringent and diverse data sets for DNA-binding proteins based on various

biologically relevant considerations: (i) seven structural classes, (ii) 86

folds, (iii) 106 superfamilies, (iv) 194 families, (v) 15 binding motifs, (vi)

single/double-stranded DNA, (vii) DNA conformation (A, B, Z, etc.), (viii) three

functions and (ix) disordered regions. These data sets were culled as

non-redundant with sequence identities of 25 and 40% and used to evaluate the

performance of 11 different methods in which online services or standalone

programs are available. We observed that the best performing methods for each of

the data sets showed significant biases toward the data sets selected for their

benchmark. Our analysis revealed important data set features, which could be used

to estimate these context-specific biases and hence suggest the best method to be

used for a given problem. We have developed a web server, which considers these

features on demand and displays the best method that the investigator should use.

The web server is freely available at http://www.biotech.iitm.ac.in/DNA-protein/.

Further, we have grouped the methods based on their complexity and analyzed the

performance. The information gained in this work could be effectively used to

select the best method for designing experiments.

DOI: 10.1093/nar/gkt544

PMCID: PMC3763535

PMID: 23788679 [Indexed for MEDLINE]

1032. OMICS. 2013 Sep;17(9):486-93. doi: 10.1089/omi.2013.0011. Epub 2013 Jun 29.

Effective classification of microRNA precursors using feature mining and AdaBoost

algorithms.

Zhong L(1), Wang JT, Wen D, Aris V, Soteropoulos P, Shapiro BA.

Author information:

(1)Bioinformatics Program and Department of Computer Science, New Jersey

Institute of Technology, Newark, New Jersey 07102, USA.

MicroRNAs play important roles in most biological processes, including cell

proliferation, tissue differentiation, and embryonic development, among others.

They originate from precursor transcripts (pre-miRNAs), which contain

phylogenetically conserved stem-loop structures. An important bioinformatics

problem is to distinguish the pre-miRNAs from pseudo pre-miRNAs that have similar

stem-loop structures. We present here a novel method for tackling this

bioinformatics problem. Our method, named MirID, accepts an RNA sequence as

input, and classifies the RNA sequence either as positive (i.e., a real

pre-miRNA) or as negative (i.e., a pseudo pre-miRNA). MirID employs a feature

mining algorithm for finding combinations of features suitable for building

pre-miRNA classification models. These models are implemented using support

vector machines, which are combined to construct a classifier ensemble. The

accuracy of the classifier ensemble is further enhanced by the utilization of an

AdaBoost algorithm. When compared with two closely related tools on twelve

species analyzed with these tools, MirID outperforms the existing tools on the

majority of the twelve species. MirID was also tested on nine additional species,

and the results showed high accuracies on the nine species. The MirID web server

is fully operational and freely accessible at

http://bioinformatics.njit.edu/MirID/ . Potential applications of this software

in genomics and medicine are also discussed.

DOI: 10.1089/omi.2013.0011

PMCID: PMC3760050

PMID: 23808606 [Indexed for MEDLINE]

1033. Proteins. 2013 Sep;81(9):1634-43. doi: 10.1002/prot.24322. Epub 2013 Jun 17.

Identification of efflux proteins using efficient radial basis function networks

with position-specific scoring matrices and biochemical properties.

Ou YY(1), Chen SA, Chang YM, Velmurugan D, Fukui K, Michael Gromiha M.

Author information:

(1)Department of Computer Science and Engineering, Yuan Ze University, Chung-Li,

Taiwan.

Efflux proteins are membrane proteins, which are involved in the transportation

of multidrugs. The annotation of efflux proteins in genomic sequences would aid

to understand the function. Although the percentage of membrane proteins in

genomes is estimated to be 25-30%, there is no information about the content of

efflux proteins. For annotating such class of proteins it is necessary to develop

a reliable method to identify efflux proteins from amino acid sequence

information. In this work, we have developed a method based on radial basis

function networks using position specific scoring matrices (PSSM) and amino acid

properties. We noticed that the C-terminal domain of efflux proteins contain

vital information for discrimination. Our method showed an accuracy of 78 and 92%

in discriminating efflux proteins from transporters and membrane proteins,

respectively using fivefold cross-validation. We utilized our method for

annotating the genomes E. coli and P. aeruginosa and it predicted 8.7 and 9.2% of

proteins as efflux proteins in these genomes, respectively. The predicted efflux

proteins have been compared with available experimental data and we observed a

very good agreement between them. Further, we developed a web server for

classifying efflux proteins and it is freely available at

http://rbf.bioinfo.tw/∼sachen/EFFLUXpredict/Efflux-RBF.php. We suggest that our

method could be an effective tool for annotating efflux proteins in genomic

sequences.

Copyright © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24322

PMID: 23670815 [Indexed for MEDLINE]

1034. PLoS One. 2013 Aug 27;8(8):e72234. doi: 10.1371/journal.pone.0072234. eCollection

2013.

iGPCR-drug: a web server for predicting interaction between GPCRs and drugs in

cellular networking.

Xiao X(1), Min JL, Wang P, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China ;

Information School, ZheJiang Textile and Fashion College, NingBo, China ; Gordon

Life Science Institute, Belmont, Massachusetts, United States of America.

Involved in many diseases such as cancer, diabetes, neurodegenerative,

inflammatory and respiratory disorders, G-protein-coupled receptors (GPCRs) are

among the most frequent targets of therapeutic drugs. It is time-consuming and

expensive to determine whether a drug and a GPCR are to interact with each other

in a cellular network purely by means of experimental techniques. Although some

computational methods were developed in this regard based on the knowledge of the

3D (dimensional) structure of protein, unfortunately their usage is quite limited

because the 3D structures for most GPCRs are still unknown. To overcome the

situation, a sequence-based classifier, called "iGPCR-drug", was developed to

predict the interactions between GPCRs and drugs in cellular networking. In the

predictor, the drug compound is formulated by a 2D (dimensional) fingerprint via

a 256D vector, GPCR by the PseAAC (pseudo amino acid composition) generated with

the grey model theory, and the prediction engine is operated by the fuzzy

K-nearest neighbour algorithm. Moreover, a user-friendly web-server for

iGPCR-drug was established at http://www.jci-bioinfo.cn/iGPCR-Drug/. For the

convenience of most experimental scientists, a step-by-step guide is provided on

how to use the web-server to get the desired results without the need to follow

the complicated math equations presented in this paper just for its integrity.

The overall success rate achieved by iGPCR-drug via the jackknife test was 85.5%,

which is remarkably higher than the rate by the existing peer method developed in

2010 although no web server was ever established for it. It is anticipated that

iGPCR-Drug may become a useful high throughput tool for both basic research and

drug development, and that the approach presented here can also be extended to

study other drug - target interaction networks.

DOI: 10.1371/journal.pone.0072234

PMCID: PMC3754978

PMID: 24015221 [Indexed for MEDLINE]

1035. JMIR Res Protoc. 2013 Aug 26;2(2):e34. doi: 10.2196/resprot.2768.

Low-intensity self-management intervention for persons with type 2 diabetes using

a mobile phone-based diabetes diary, with and without health counseling and

motivational interviewing: protocol for a randomized controlled trial.

Ribu L(1), Holmen H, Torbjørnsen A, Wahl AK, Grøttland A, Småstuen MC, Elind E,

Bergmo TS, Breivik E, Arsand E.

Author information:

(1)Department of Nursing, Faculty of Health Sciences, Oslo and Akershus

University College of Applied Sciences, Oslo, Norway. Lis.Ribu@hioa.no.

BACKGROUND: The present study protocol is designed to cover the Norwegian part of

the European Union Collaborative Project-REgioNs of Europe WorkINg together for

HEALTH (RENEWING HEALTH). Self-management support is an important element of care

for persons with type 2 diabetes (T2D) for achieving metabolic control and

positive lifestyle changes. Telemedicine (TM) with or without health counseling

may become an important technological aid for self-management and may provide a

user-centered model of care. In spite of many earlier studies on TM, there

remains a lack of consensus in research findings about the effect of TM

interventions.

OBJECTIVE: The aim of RENEWING HEALTH is to validate and evaluate innovative TM

tools on a large scale through a common evaluation, making it easier for decision

makers to choose the most efficient and cost-effective technological

interventions. The Norwegian pilot study evaluates whether the introduction of a

mobile phone with a diabetes diary application together with health counseling

intervention produces benefits in terms of the desired outcomes, as reflected in

the hemoglobin A1c level, health-related quality of life, behavior change, and

cost-effectiveness.

METHODS: The present study has a mixed-method design comprising a three-armed

prospective randomized controlled trial and qualitative interviews with study

data collected at three time points: baseline, after 4 months, and after 1 year.

The patients' registrations on the application are recorded continuously and are

sent securely to a server.

RESULTS: The inclusion of patients started in March 2011, and 100% of the planned

sample size is included (N=151). Of all the participants, 26/151 patients (17.2%)

are lost to follow-up by now, and 11/151 patients (7.3%) are still in the trial.

Results of the study protocol will be presented in 2014.

CONCLUSIONS: The key goals of this trial are to investigate the effect of an

electronic diabetes diary app with and without health counseling, and to

determine whether health counseling is important to the continued use of the

application and the patients' health competence and acceptability. Research

within this area is needed because few studies have investigated the

effectiveness of apps used in long-term interventions with this degree of

self-management.

TRIAL REGISTRATION: Clinicaltrials.gov NCT01315756;

http://clinicaltrials.gov/ct2/show/NCT01315756 (Archived by WebCite at

http://www.webcitation/6BTyuRMpH).

DOI: 10.2196/resprot.2768

PMCID: PMC3758066

PMID: 23978690

1036. J Med Internet Res. 2013 Aug 19;15(8):e170. doi: 10.2196/jmir.2566.

Usage of a generic web-based self-management intervention for breast cancer

survivors: substudy analysis of the BREATH trial.

van den Berg SW(1), Peters EJ, Kraaijeveld JF, Gielissen MF, Prins JB.

Author information:

(1)Department of Medical Psychology, Radboud University Nijmegen Medical Centre,

Nijmegen, Netherlands. s.vandenberg@mps.umcn.nl

BACKGROUND: Generic fully automated Web-based self-management interventions are

upcoming, for example, for the growing number of breast cancer survivors. It is

hypothesized that the use of these interventions is more individualized and that

users apply a large amount of self-tailoring. However, technical usage

evaluations of these types of interventions are scarce and practical guidelines

are lacking.

OBJECTIVE: To gain insight into meaningful usage parameters to evaluate the use

of generic fully automated Web-based interventions by assessing how breast cancer

survivors use a generic self-management website. Final aim is to propose

practical recommendations for researchers and information and communication

technology (ICT) professionals who aim to design and evaluate the use of similar

Web-based interventions.

METHODS: The BREAst cancer ehealTH (BREATH) intervention is a generic unguided

fully automated website with stepwise weekly access and a fixed 4-month structure

containing 104 intervention ingredients (ie, texts, tasks, tests, videos). By

monitoring https-server requests, technical usage statistics were recorded for

the intervention group of the randomized controlled trial. Observed usage was

analyzed by measures of frequency, duration, and activity. Intervention adherence

was defined as continuous usage, or the proportion of participants who started

using the intervention and continued to log in during all four phases. By

comparing observed to minimal intended usage (frequency and activity), different

user groups were defined.

RESULTS: Usage statistics for 4 months were collected from 70 breast cancer

survivors (mean age 50.9 years). Frequency of logins/person ranged from 0 to 45,

total duration/person from 0 to 2324 minutes (38.7 hours), and activity from

opening none to all intervention ingredients. 31 participants continued logging

in to all four phases resulting in an intervention adherence rate of 44.3% (95%

CI 33.2-55.9). Nine nonusers (13%), 30 low users (43%), and 31 high users (44%)

were defined. Low and high users differed significantly on frequency (P<.001),

total duration (P<.001), session duration (P=.009), and activity (P<.001). High

users logged in an average of 21 times, had a mean session duration of 33

minutes, and opened on average 91% of all ingredients. Signing the self-help

contract (P<.001), reporting usefulness of ingredients (P=.003), overall

satisfaction (P=.028), and user friendliness evaluation (P=.003) were higher in

high users. User groups did not differ on age, education, and baseline distress.

CONCLUSIONS: By reporting the usage of a self-management website for breast

cancer survivors, the present study gained first insight into the design of usage

evaluations of generic fully automated Web-based interventions. It is recommended

to (1) incorporate usage statistics that reflect the amount of self-tailoring

applied by users, (2) combine technical usage statistics with self-reported

usefulness, and (3) use qualitative measures. Also, (4) a pilot usage evaluation

should be a fixed step in the development process of novel Web-based

interventions, and (5) it is essential for researchers to gain insight into the

rationale of recorded and nonrecorded usage statistics.

TRIAL REGISTRATION: Netherlands Trial Register (NTR): 2935;

http://www.trialregister.nl/trialreg/admin/rctview.asp?TC=2935 (Archived by

WebCite at http://www.webcitation.org/6IkX1ADEV).

DOI: 10.2196/jmir.2566

PMCID: PMC3758022

PMID: 23958584 [Indexed for MEDLINE]

1037. BMC Bioinformatics. 2013 Aug 16;14:249. doi: 10.1186/1471-2105-14-249.

Gentrepid V2.0: a web server for candidate disease gene prediction.

Ballouz S(1), Liu JY, George RA, Bains N, Liu A, Oti M, Gaeta B, Fatkin D,

Wouters MA.

Author information:

(1)School of Medicine, Deakin University, Geelong, VIC 3217, Australia.

m.wouters@deakin.edu.au.

BACKGROUND: Candidate disease gene prediction is a rapidly developing area of

bioinformatics research with the potential to deliver great benefits to human

health. As experimental studies detecting associations between genetic intervals

and disease proliferate, better bioinformatic techniques that can expand and

exploit the data are required.

DESCRIPTION: Gentrepid is a web resource which predicts and prioritizes candidate

disease genes for both Mendelian and complex diseases. The system can take input

from linkage analysis of single genetic intervals or multiple marker loci from

genome-wide association studies. The underlying database of the Gentrepid tool

sources data from numerous gene and protein resources, taking advantage of the

wealth of biological information available. Using known disease gene information

from OMIM, the system predicts and prioritizes disease gene candidates that

participate in the same protein pathways or share similar protein domains.

Alternatively, using an ab initio approach, the system can detect enrichment of

these protein annotations without prior knowledge of the phenotype.

CONCLUSIONS: The system aims to integrate the wealth of protein information

currently available with known and novel phenotype/genotype information to

acquire knowledge of biological mechanisms underpinning disease. We have updated

the system to facilitate analysis of GWAS data and the study of complex diseases.

Application of the system to GWAS data on hypertension using the ICBP data is

provided as an example. An interesting prediction is a ZIP transporter additional

to the one found by the ICBP analysis. The webserver URL is

https://www.gentrepid.org/.

DOI: 10.1186/1471-2105-14-249

PMCID: PMC3844418

PMID: 23947436 [Indexed for MEDLINE]

1038. Bioinformatics. 2013 Aug 15;29(16):2051-2. doi: 10.1093/bioinformatics/btt325.

Epub 2013 Jun 5.

WhichCyp: prediction of cytochromes P450 inhibition.

Rostkowski M(1), Spjuth O, Rydberg P.

Author information:

(1)Department of Drug Design and Pharmacology, University of Copenhagen,

Universitetsparken 2, Copenhagen, Denmark.

SUMMARY: In this work we present WhichCyp, a tool for prediction of which

cytochromes P450 isoforms (among 1A2, 2C9, 2C19, 2D6 and 3A4) a given molecule is

likely to inhibit. The models are built from experimental high-throughput data

using support vector machines and molecular signatures.

AVAILABILITY: The WhichCyp server is freely available for use on the web at

http://drug.ku.dk/whichcyp, where the WhichCyp Java program and source code is

also available for download.

DOI: 10.1093/bioinformatics/btt325

PMID: 23740742 [Indexed for MEDLINE]

1039. PLoS One. 2013 Aug 15;8(8):e72343. doi: 10.1371/journal.pone.0072343. eCollection

2013.

Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based

on orthology and phylogeny.

Wei W(1), Ning LW, Ye YN, Guo FB.

Author information:

(1)Center of Bioinformatics and Key Laboratory for NeuroInformation of Ministry

of Education, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu, China.

Integrative genomics predictors, which score highly in predicting bacterial

essential genes, would be unfeasible in most species because the data sources are

limited. We developed a universal approach and tool designated Geptop, based on

orthology and phylogeny, to offer gene essentiality annotations. In a series of

tests, our Geptop method yielded higher area under curve (AUC) scores in the

receiver operating curves than the integrative approaches. In the ten-fold

cross-validations among randomly upset samples, Geptop yielded an AUC of 0.918,

and in the cross-organism predictions for 19 organisms Geptop yielded AUC scores

between 0.569 and 0.959. A test applied to the very recently determined essential

gene dataset from the Porphyromonas gingivalis, which belongs to a phylum

different with all of the above 19 bacterial genomes, gave an AUC of 0.77.

Therefore, Geptop can be applied to any bacterial species whose genome has been

sequenced. Compared with the essential genes uniquely identified by the lethal

screening, the essential genes predicted only by Gepop are associated with more

protein-protein interactions, especially in the three bacteria with lower AUC

scores (<0.7). This may further illustrate the reliability and feasibility of our

method in some sense. The web server and standalone version of Geptop are

available at http://cefg.uestc.edu.cn/geptop/ free of charge. The tool has been

run on 968 bacterial genomes and the results are accessible at the website.

DOI: 10.1371/journal.pone.0072343

PMCID: PMC3744497

PMID: 23977285 [Indexed for MEDLINE]

1040. BMC Bioinformatics. 2013 Aug 13;14:247. doi: 10.1186/1471-2105-14-247.

PKIS: computational identification of protein kinases for experimentally

discovered protein phosphorylation sites.

Zou L(1), Wang M, Shen Y, Liao J, Li A, Wang M.

Author information:

(1)Department of Electronic Science and Technology, University of Science and

Technology of China, Hefei 230027, China. mhwang@ustc.edu.cn.

BACKGROUND: Dynamic protein phosphorylation is an essential regulatory mechanism

in various organisms. In this capacity, it is involved in a multitude of signal

transduction pathways. Kinase-specific phosphorylation data lay the foundation

for reconstruction of signal transduction networks. For this reason, precise

annotation of phosphorylated proteins is the first step toward simulating cell

signaling pathways. However, the vast majority of kinase-specific phosphorylation

data remain undiscovered and existing experimental methods and computational

phosphorylation site (P-site) prediction tools have various limitations with

respect to addressing this problem.

RESULTS: To address this issue, a novel protein kinase identification web server,

PKIS, is here presented for the identification of the protein kinases responsible

for experimentally verified P-sites at high specificity, which incorporates the

composition of monomer spectrum (CMS) encoding strategy and support vector

machines (SVMs). Compared to widely used P-site prediction tools including

KinasePhos 2.0, Musite, and GPS2.1, PKIS largely outperformed these tools in

identifying protein kinases associated with known P-sites. In addition, PKIS was

used on all the P-sites in Phospho.ELM that currently lack kinase information. It

successfully identified 14 potential SYK substrates with 36 known P-sites.

Further literature search showed that 5 of them were indeed phosphorylated by

SYK. Finally, an enrichment analysis was performed and 6 significant SYK-related

signal pathways were identified.

CONCLUSIONS: In general, PKIS can identify protein kinases for experimental

phosphorylation sites efficiently. It is a valuable bioinformatics tool suitable

for the study of protein phosphorylation. The PKIS web server is freely available

at http://bioinformatics.ustc.edu.cn/pkis.

DOI: 10.1186/1471-2105-14-247

PMCID: PMC3765618

PMID: 23941207 [Indexed for MEDLINE]

1041. PLoS One. 2013 Aug 5;8(8):e70151. doi: 10.1371/journal.pone.0070151. Print 2013.

Filtering for compound heterozygous sequence variants in non-consanguineous

pedigrees.

Kamphans T(1), Sabri P, Zhu N, Heinrich V, Mundlos S, Robinson PN, Parkhomchuk D,

Krawitz PM.

Author information:

(1)Smart Algos, Berlin, Germany.

The identification of disease-causing mutations in next-generation sequencing

(NGS) data requires efficient filtering techniques. In patients with rare

recessive diseases, compound heterozygosity of pathogenic mutations is the most

likely inheritance model if the parents are non-consanguineous. We developed a

web-based compound heterozygous filter that is suited for data from NGS projects

and that is easy to use for non-bioinformaticians. We analyzed the power of

compound heterozygous mutation filtering by deriving background distributions for

healthy individuals from different ethnicities and studied the effectiveness in

trios as well as more complex pedigree structures. While usually more then 30

genes harbor potential compound heterozygotes in single exomes, this number can

be markedly reduced with every additional member of the pedigree that is included

in the analysis. In a real data set with exomes of four family members, two

sisters affected by Mabry syndrome and their healthy parents, the disease-causing

gene PIGO, which harbors the pathogenic compound heterozygous variants, could be

readily identified. Compound heterozygous filtering is an efficient means to

reduce the number of candidate mutations in studies aiming at identifying

recessive disease genes in non-consanguineous families. A web-server is provided

to make this filtering strategy available at www.gene-talk.de.

DOI: 10.1371/journal.pone.0070151

PMCID: PMC3734130

PMID: 23940540 [Indexed for MEDLINE]

1042. Biochim Biophys Acta. 2013 Aug;1834(8):1671-80. doi:

10.1016/j.bbapap.2013.05.022. Epub 2013 Jun 1.

RAPID: fast and accurate sequence-based prediction of intrinsic disorder content

on proteomic scale.

Yan J(1), Mizianty MJ, Filipow PL, Uversky VN, Kurgan L.

Author information:

(1)Department of Electrical and Computer Engineering, University of Alberta,

Edmonton, Canada.

Recent research in the protein intrinsic disorder was stimulated by the

availability of accurate computational predictors. However, most of these methods

are relatively slow, especially considering proteome-scale applications, and were

shown to produce relatively large errors when estimating disorder at the protein-

(in contrast to residue-) level, which is defined by the fraction/content of

disordered residues. To this end, we propose a novel support vector

Regression-based Accurate Predictor of Intrinsic Disorder (RAPID). Key advantages

of RAPID are speed (prediction of an average-size eukaryotic proteome takes <1h

on a modern desktop computer); sophisticated design (multiple, complementary

information sources that are aggregated over an input chain are combined using

feature selection); and high-quality and robust predictive performance. Empirical

tests on two diverse benchmark datasets reveal that RAPID's predictive

performance compares favorably to a comprehensive set of state-of-the-art

disorder and disorder content predictors. Drawing on high speed and good

predictive quality, RAPID was used to perform large-scale characterization of

disorder in 200+ fully sequenced eukaryotic proteomes. Our analysis reveals

interesting relations of disorder with structural coverage and chain length, and

unusual distribution of fully disordered chains. We also performed a

comprehensive (using 56000+ annotated chains, which doubles the scope of previous

studies) investigation of cellular functions and localizations that are enriched

in the disorder in the human proteome. RAPID, which allows for batch

(proteome-wide) predictions, is available as a web server at

http://biomine.ece.ualberta.ca/RAPID/.

Copyright © 2013 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.bbapap.2013.05.022

PMID: 23732563 [Indexed for MEDLINE]

1043. Biochim Biophys Acta. 2013 Aug;1834(8):1461-7. doi: 10.1016/j.bbapap.2013.04.006.

Epub 2013 Apr 19.

hCKSAAP\_UbSite: improved prediction of human ubiquitination sites by exploiting

amino acid pattern and properties.

Chen Z(1), Zhou Y, Song J, Zhang Z.

Author information:

(1)College of Biological Sciences, China Agricultural University, Beijing, China.

As one of the most common post-translational modifications, ubiquitination

regulates the quantity and function of a variety of proteins. Experimental and

clinical investigations have also suggested the crucial roles of ubiquitination

in several human diseases. The complicated sequence context of human

ubiquitination sites revealed by proteomic studies highlights the need of

developing effective computational strategies to predict human ubiquitination

sites. Here we report the establishment of a novel human-specific ubiquitination

site predictor through the integration of multiple complementary classifiers.

Firstly, a Support Vector Machine (SVM) classier was constructed based on the

composition of k-spaced amino acid pairs (CKSAAP) encoding, which has been

utilized in our previous yeast ubiquitination site predictor. To further exploit

the pattern and properties of the ubiquitination sites and their flanking

residues, three additional SVM classifiers were constructed using the binary

amino acid encoding, the AAindex physicochemical property encoding and the

protein aggregation propensity encoding, respectively. Through an integration

that relied on logistic regression, the resulting predictor termed hCKSAAP\_UbSite

achieved an area under ROC curve (AUC) of 0.770 in 5-fold cross-validation test

on a class-balanced training dataset. When tested on a class-balanced independent

testing dataset that contains 3419 ubiquitination sites, hCKSAAP\_UbSite has also

achieved a robust performance with an AUC of 0.757. Specifically, it has

consistently performed better than the predictor using the CKSAAP encoding alone

and two other publicly available predictors which are not human-specific. Given

its promising performance in our large-scale datasets, hCKSAAP\_UbSite has been

made publicly available at our server (http://protein.cau.edu.cn/cksaap\_ubsite/).

Copyright © 2013 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.bbapap.2013.04.006

PMID: 23603789 [Indexed for MEDLINE]

1044. Bioinformatics. 2013 Aug 1;29(15):1922-4. doi: 10.1093/bioinformatics/btt316.

Epub 2013 Jun 3.

Relating genes to function: identifying enriched transcription factors using the

ENCODE ChIP-Seq significance tool.

Auerbach RK(1), Chen B, Butte AJ.

Author information:

(1)Division of Systems Medicine, Department of Pediatrics, Stanford University

School of Medicine, 1265 Welch Road, Room X-163 MS-5415, Stanford, CA 94305, USA.

MOTIVATION: Biological analysis has shifted from identifying genes and

transcripts to mapping these genes and transcripts to biological functions. The

ENCODE Project has generated hundreds of ChIP-Seq experiments spanning multiple

transcription factors and cell lines for public use, but tools for a biomedical

scientist to analyze these data are either non-existent or tailored to narrow

biological questions. We present the ENCODE ChIP-Seq Significance Tool, a

flexible web application leveraging public ENCODE data to identify enriched

transcription factors in a gene or transcript list for comparative analyses.

IMPLEMENTATION: The ENCODE ChIP-Seq Significance Tool is written in JavaScript on

the client side and has been tested on Google Chrome, Apple Safari and Mozilla

Firefox browsers. Server-side scripts are written in PHP and leverage R and a

MySQL database. The tool is available at http://encodeqt.stanford.edu.

CONTACT: abutte@stanford.edu

SUPPLEMENTARY INFORMATION: Supplementary material is available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt316

PMCID: PMC3712221

PMID: 23732275 [Indexed for MEDLINE]

1045. Bioinformatics. 2013 Aug 1;29(15):1910-2. doi: 10.1093/bioinformatics/btt303.

Epub 2013 May 28.

HitPick: a web server for hit identification and target prediction of chemical

screenings.

Liu X(1), Vogt I, Haque T, Campillos M.

Author information:

(1)Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München,

85764 Neuherberg, Germany.

MOTIVATION: High-throughput phenotypic assays reveal information about the

molecules that modulate biological processes, such as a disease phenotype and a

signaling pathway. In these assays, the identification of hits along with their

molecular targets is critical to understand the chemical activities modulating

the biological system. Here, we present HitPick, a web server for identification

of hits in high-throughput chemical screenings and prediction of their molecular

targets. HitPick applies the B-score method for hit identification and a newly

developed approach combining 1-nearest-neighbor (1NN) similarity searching and

Laplacian-modified naïve Bayesian target models to predict targets of identified

hits. The performance of the HitPick web server is presented and discussed.

AVAILABILITY: The server can be accessed at

http://mips.helmholtz-muenchen.de/proj/hitpick.

CONTACT: monica.campillos@helmholtz-muenchen.de.

DOI: 10.1093/bioinformatics/btt303

PMID: 23716196 [Indexed for MEDLINE]

1046. Nucleic Acids Res. 2013 Aug;41(14):e138. doi: 10.1093/nar/gkt435. Epub 2013 May

22.

CLIP-based prediction of mammalian microRNA binding sites.

Liu C(1), Mallick B, Long D, Rennie WA, Wolenc A, Carmack CS, Ding Y.

Author information:

(1)Wadsworth Center, New York State Department of Health, Center for Medical

Science, 150 New Scotland Avenue, Albany, NY 12208, USA.

Prediction and validation of microRNA (miRNA) targets are essential for

understanding functions of miRNAs in gene regulation. Crosslinking

immunoprecipitation (CLIP) allows direct identification of a huge number of

Argonaute-bound target sequences that contain miRNA binding sites. By analysing

data from CLIP studies, we identified a comprehensive list of sequence,

thermodynamic and target structure features that are essential for target binding

by miRNAs in the 3' untranslated region (3' UTR), coding sequence (CDS) region

and 5' untranslated region (5' UTR) of target messenger RNA (mRNA). The total

energy of miRNA:target hybridization, a measure of target structural

accessibility, is the only essential feature common for both seed and seedless

sites in all three target regions. Furthermore, evolutionary conservation is an

important discriminating feature for both seed and seedless sites. These features

enabled us to develop novel statistical models for the predictions of both seed

sites and broad classes of seedless sites. Through both intra-dataset validation

and inter-dataset validation, our approach showed major improvements over

established algorithms for predicting seed sites and a class of seedless sites.

Furthermore, we observed good performance from cross-species validation,

suggesting that our prediction framework can be valuable for broad application to

other mammalian species and beyond. Transcriptome-wide binding site predictions

enabled by our approach will greatly complement the available CLIP data, which

only cover small fractions of transcriptomes and known miRNAs due to

non-detectable levels of expression. Software and database tools based on the

prediction models have been developed and are available through Sfold web server

at http://sfold.wadsworth.org.

DOI: 10.1093/nar/gkt435

PMCID: PMC3737542

PMID: 23703212 [Indexed for MEDLINE]

1047. PLoS One. 2013 Jul 25;8(7):e68370. doi: 10.1371/journal.pone.0068370. Print 2013.

Prediction of disease causing non-synonymous SNPs by the Artificial Neural

Network Predictor NetDiseaseSNP.

Johansen MB(1), Izarzugaza JM, Brunak S, Petersen TN, Gupta R.

Author information:

(1)Center for Biological Sequence Analysis, Department of Systems Biology,

Technical University of Denmark, Kongens Lyngby, Denmark.

We have developed a sequence conservation-based artificial neural network

predictor called NetDiseaseSNP which classifies nsSNPs as disease-causing or

neutral. Our method uses the excellent alignment generation algorithm of SIFT to

identify related sequences and a combination of 31 features assessing sequence

conservation and the predicted surface accessibility to produce a single score

which can be used to rank nsSNPs based on their potential to cause disease.

NetDiseaseSNP classifies successfully disease-causing and neutral mutations. In

addition, we show that NetDiseaseSNP discriminates cancer driver and passenger

mutations satisfactorily. Our method outperforms other state-of-the-art methods

on several disease/neutral datasets as well as on cancer driver/passenger

mutation datasets and can thus be used to pinpoint and prioritize plausible

disease candidates among nsSNPs for further investigation. NetDiseaseSNP is

publicly available as an online tool as well as a web service:

http://www.cbs.dtu.dk/services/NetDiseaseSNP.

DOI: 10.1371/journal.pone.0068370

PMCID: PMC3723835

PMID: 23935863 [Indexed for MEDLINE]

1048. PLoS One. 2013 Jul 19;8(7):e69648. doi: 10.1371/journal.pone.0069648. Print 2013.

i3Drefine software for protein 3D structure refinement and its assessment in

CASP10.

Bhattacharya D(1), Cheng J.

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, Missouri,

United States of America.

Protein structure refinement refers to the process of improving the qualities of

protein structures during structure modeling processes to bring them closer to

their native states. Structure refinement has been drawing increasing attention

in the community-wide Critical Assessment of techniques for Protein Structure

prediction (CASP) experiments since its addition in 8(th) CASP experiment. During

the 9(th) and recently concluded 10(th) CASP experiments, a consistent growth in

number of refinement targets and participating groups has been witnessed. Yet,

protein structure refinement still remains a largely unsolved problem with

majority of participating groups in CASP refinement category failed to

consistently improve the quality of structures issued for refinement. In order to

alleviate this need, we developed a completely automated and computationally

efficient protein 3D structure refinement method, i3Drefine, based on an

iterative and highly convergent energy minimization algorithm with a powerful

all-atom composite physics and knowledge-based force fields and hydrogen bonding

(HB) network optimization technique. In the recent community-wide blind

experiment, CASP10, i3Drefine (as 'MULTICOM-CONSTRUCT') was ranked as the best

method in the server section as per the official assessment of CASP10 experiment.

Here we provide the community with free access to i3Drefine software and

systematically analyse the performance of i3Drefine in strict blind mode on the

refinement targets issued in CASP10 refinement category and compare with other

state-of-the-art refinement methods participating in CASP10. Our analysis

demonstrates that i3Drefine is only fully-automated server participating in

CASP10 exhibiting consistent improvement over the initial structures in both

global and local structural quality metrics. Executable version of i3Drefine is

freely available at http://protein.rnet.missouri.edu/i3drefine/.

DOI: 10.1371/journal.pone.0069648

PMCID: PMC3716612

PMID: 23894517 [Indexed for MEDLINE]

1049. Bioinformatics. 2013 Jul 15;29(14):1811-2. doi: 10.1093/bioinformatics/btt283.

Epub 2013 May 29.

COPRED: prediction of fold, GO molecular function and functional residues at the

domain level.

López D(1), Pazos F.

Author information:

(1)Systems Biology Department, Computational Systems Biology Group CNB-CSIC, c/

Darwin, 3. Cantoblanco, 28049 Madrid, Spain.

SUMMARY: Only recently the first resources devoted to the functional annotation

of proteins at the domain level started to appear. The next step is to develop

specific methodologies for predicting function at the domain level based on these

resources, and to implement them in web servers to be used by the community. In

this work, we present COPRED, a web server for the concomitant prediction of

fold, molecular function and functional sites at the domain level, based on a

methodology for domain molecular function prediction and a resource of domain

functional annotations previously developed and benchmarked.

AVAILABILITY AND IMPLEMENTATION: COPRED can be freely accessed at

http://csbg.cnb.csic.es/copred. The interface works in all standard web browsers.

WebGL (natively supported by most browsers) is required for the in-line preview

and manipulation of protein 3D structures. The website includes a detailed help

section and usage examples.

CONTACT: pazos@cnb.csic.es.

DOI: 10.1093/bioinformatics/btt283

PMID: 23720488 [Indexed for MEDLINE]

1050. Bioinformatics. 2013 Jul 15;29(14):1827-9. doi: 10.1093/bioinformatics/btt270.

Epub 2013 May 27.

ChemMapper: a versatile web server for exploring pharmacology and chemical

structure association based on molecular 3D similarity method.

Gong J(1), Cai C, Liu X, Ku X, Jiang H, Gao D, Li H.

Author information:

(1)School of Information Science and Engineering, East China University of

Science and Technology, Shanghai 200237, China.

SUMMARY: ChemMapper is an online platform to predict polypharmacology effect and

mode of action for small molecules based on 3D similarity computation. ChemMapper

collects >350 000 chemical structures with bioactivities and associated target

annotations (as well as >3 000 000 non-annotated compounds for virtual

screening). Taking the user-provided chemical structure as the query, the top

most similar compounds in terms of 3D similarity are returned with associated

pharmacology annotations. ChemMapper is designed to provide versatile services in

a variety of chemogenomics, drug repurposing, polypharmacology, novel bioactive

compounds identification and scaffold hopping studies.

AVAILABILITY: http://lilab.ecust.edu.cn/chemmapper/.

CONTACT: xfliu@ecust.edu.cn or hlli@ecust.edu.cn

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt270

PMID: 23712658 [Indexed for MEDLINE]

1051. Bioinformatics. 2013 Jul 15;29(14):1750-7. doi: 10.1093/bioinformatics/btt278.

Epub 2013 May 15.

CAPITO--a web server-based analysis and plotting tool for circular dichroism

data.

Wiedemann C(1), Bellstedt P, Görlach M.

Author information:

(1)Biomolecular NMR Spectroscopy, Leibniz Institute for Age Research-Fritz

Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany. cwiede@fli-leibniz.de

MOTIVATION: Circular dichroism (CD) spectroscopy is one of the most versatile

tools to study protein folding and to validate the proper fold of purified

proteins. Here, we aim to provide a readily accessible, user-friendly and

platform-independent tool capable of analysing multiple CD datasets of virtually

any format and returning results as high-quality graphical output to the user.

RESULTS: CAPITO (CD Anaylsis and Plotting Tool) is a novel web server-based tool

for analysing and plotting CD data. It allows reliable estimation of secondary

structure content utilizing different approaches. CAPITO accepts multiple CD

datasets and, hence, is well suited for a wide application range such as the

analysis of temperature or pH-dependent (un)folding and the comparison of

mutants.

AVAILABILITY: http://capito.nmr.fli-leibniz.de.

CONTACT: cwiede@fli-leibniz.de or mago@fli-leibniz.de

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt278

PMID: 23681122 [Indexed for MEDLINE]

1052. Bioinformatics. 2013 Jul 15;29(14):1834-6. doi: 10.1093/bioinformatics/btt279.

Epub 2013 May 15.

DrugMap Central: an on-line query and visualization tool to facilitate drug

repositioning studies.

Fu C(1), Jin G, Gao J, Zhu R, Ballesteros-Villagrana E, Wong ST.

Author information:

(1)Department of Systems Medicine and Bioengineering, The Methodist Hospital

Research Institute, Weill Cornell Medical College.

SUMMARY: Systematic studies of drug repositioning require the integration of

multi-level drug data, including basic chemical information (such as SMILES),

drug targets, target-related signaling pathways, clinical trial information and

Food and Drug Administration (FDA)-approval information, to predict new potential

indications of existing drugs. Currently available databases, however, lack query

support for multi-level drug information and thus are not designed to support

drug repositioning studies. DrugMap Central (DMC), an online tool, is developed

to help fill the gap. DMC enables the users to integrate, query, visualize,

interrogate, and download multi-level data of known drugs or compounds quickly

for drug repositioning studies all within one system.

AVAILABILITY: DMC is accessible at http://r2d2drug.org/DMC.aspx.

CONTACT: STWong@tmhs.org.

DOI: 10.1093/bioinformatics/btt279

PMCID: PMC3702253

PMID: 23681121 [Indexed for MEDLINE]

1053. Algorithms Mol Biol. 2013 Jul 11;8(1):19. doi: 10.1186/1748-7188-8-19.

MORPH-PRO: a novel algorithm and web server for protein morphing.

Castellana NE(1), Lushnikov A, Rotkiewicz P, Sefcovic N, Pevzner PA, Godzik A,

Vyatkina K.

Author information:

(1)Algorithmic Biology Laboratory, Saint Petersburg Academic University, Saint

Petersburg, Russia. kira@math.spbu.ru.

BACKGROUND: Proteins are known to be dynamic in nature, changing from one

conformation to another while performing vital cellular tasks. It is important to

understand these movements in order to better understand protein function. At the

same time, experimental techniques provide us with only single snapshots of the

whole ensemble of available conformations. Computational protein morphing

provides a visualization of a protein structure transitioning from one

conformation to another by producing a series of intermediate conformations.

RESULTS: We present a novel, efficient morphing algorithm, Morph-Pro based on

linear interpolation. We also show that apart from visualization, morphing can be

used to provide plausible intermediate structures. We test this by using the

intermediate structures of a c-Jun N-terminal kinase (JNK1) conformational change

in a virtual docking experiment. The structures are shown to dock with higher

score to known JNK1-binding ligands than structures solved using X-Ray

crystallography. This experiment demonstrates the potential applications of the

intermediate structures in modeling or virtual screening efforts.

CONCLUSIONS: Visualization of protein conformational changes is important for

characterization of protein function. Furthermore, the intermediate structures

produced by our algorithm are good approximations to true structures. We believe

there is great potential for these computationally predicted structures in

protein-ligand docking experiments and virtual screening. The Morph-Pro web

server can be accessed at http://morph-pro.bioinf.spbau.ru.

DOI: 10.1186/1748-7188-8-19

PMCID: PMC3738870

PMID: 23844614

1054. Database (Oxford). 2013 Jul 11;2013:bat051. doi: 10.1093/database/bat051. Print

2013.

JBioWH: an open-source Java framework for bioinformatics data integration.

Vera R(1), Perez-Riverol Y, Perez S, Ligeti B, Kertész-Farkas A, Pongor S.

Author information:

(1)Department of Physics, Polytechnic University Jose E. Echeverria, Havana,

Cuba. Roberto.Vera@icgeb.org

The Java BioWareHouse (JBioWH) project is an open-source platform-independent

programming framework that allows a user to build his/her own integrated database

from the most popular data sources. JBioWH can be used for intensive querying of

multiple data sources and the creation of streamlined task-specific data sets on

local PCs. JBioWH is based on a MySQL relational database scheme and includes

JAVA API parser functions for retrieving data from 20 public databases (e.g.

NCBI, KEGG, etc.). It also includes a client desktop application for

(non-programmer) users to query data. In addition, JBioWH can be tailored for use

in specific circumstances, including the handling of massive queries for

high-throughput analyses or CPU intensive calculations. The framework is provided

with complete documentation and application examples and it can be downloaded

from the Project Web site at http://code.google.com/p/jbiowh. A MySQL server is

available for demonstration purposes at hydrax.icgeb.trieste.it:3307. Database

URL: http://code.google.com/p/jbiowh.

DOI: 10.1093/database/bat051

PMCID: PMC3708619

PMID: 23846595 [Indexed for MEDLINE]

1055. Front Genet. 2013 Jul 11;4:133. doi: 10.3389/fgene.2013.00133. eCollection 2013.

MicroRNA discovery by similarity search to a database of RNA-seq profiles.

Pundhir S(1), Gorodkin J.

Author information:

(1)Center for non-coding RNA in Technology and Health, Department of Veterinary

Clinical and Animal Sciences (IKVH), University of Copenhagen Frederiksberg C,

Denmark.

In silico generated search for microRNAs (miRNAs) has been driven by methods

compiling structural features of the miRNA precursor hairpin, as well as to some

degree combining this with the analysis of RNA-seq profiles for which the miRNA

typically leave the drosha/dicer fingerprint of 1-2 ~22 nt blocks of reads

corresponding to the mature and star miRNA. In complement to the previous

methods, we present a study where we systematically exploit these patterns of

read profiles. We created two datasets comprised of 2540 and 4795 read profiles

obtained after preprocessing short RNA-seq data from miRBase and ENCODE,

respectively. Out of 4795 ENCODE read profiles, 1361 are annotated as non-coding

RNAs (ncRNAs) and of which 285 are further annotated as miRNAs. Using

deepBlockAlign (dba), we align ncRNA read profiles from ENCODE against the

miRBase read profiles (cleaned for "self-matches") and are able to separate

ENCODE miRNAs from the other ncRNAs by a Matthews Correlation Coefficient (MCC)

of 0.8 and obtain an area under the curve of 0.93. Based on the dba score cut-off

of 0.7 at which we observed the maximum MCC of 0.8, we predict 523 novel miRNA

candidates. An additional RNA secondary structure analysis reveal that 42 of the

candidates overlap with predicted conserved secondary structure. Further analysis

reveal that the 523 miRNA candidates are located in genomic regions with MAF

block (UCSC) fragmentation and poor sequence conservation, which in part might

explain why they have been overlooked in previous efforts. We further analyzed

known human and mouse miRNA read profiles and found two distinct classes; the

first containing two blocks and the second containing >2 blocks of reads. Also

the latter class holds read profiles that have less well defined arrangement of

reads in comparison to the former class. On comparison of miRNA read profiles

from plants and animals, we observed kingdom specific read profiles that are

distinct in terms of both length and distribution of reads within the read

profiles to each other. All the data, as well as a server to search miRBase read

profiles by uploading a BED file, is available at http://rth.dk/resources/mirdba.

DOI: 10.3389/fgene.2013.00133

PMCID: PMC3708161

PMID: 23874353

1056. J Biotechnol. 2013 Jul 10;166(3):122-34. doi: 10.1016/j.jbiotec.2013.04.004. Epub

2013 Apr 16.

Transcriptome analysis based on next-generation sequencing of non-model plants

producing specialized metabolites of biotechnological interest.

Xiao M(1), Zhang Y, Chen X, Lee EJ, Barber CJ, Chakrabarty R, Desgagné-Penix I,

Haslam TM, Kim YB, Liu E, MacNevin G, Masada-Atsumi S, Reed DW, Stout JM, Zerbe

P, Zhang Y, Bohlmann J, Covello PS, De Luca V, Page JE, Ro DK, Martin VJ,

Facchini PJ, Sensen CW.

Author information:

(1)Department of Biochemistry and Molecular Biology, University of Calgary, 3330

Hospital Drive NW, Calgary, Alberta T2N 4N1, Canada.

Plants produce a vast array of specialized metabolites, many of which are used as

pharmaceuticals, flavors, fragrances, and other high-value fine chemicals.

However, most of these compounds occur in non-model plants for which genomic

sequence information is not yet available. The production of a large amount of

nucleotide sequence data using next-generation technologies is now relatively

fast and cost-effective, especially when using the latest Roche-454 and Illumina

sequencers with enhanced base-calling accuracy. To investigate specialized

metabolite biosynthesis in non-model plants we have established a data-mining

framework, employing next-generation sequencing and computational algorithms, to

construct and analyze the transcriptomes of 75 non-model plants that produce

compounds of interest for biotechnological applications. After sequence assembly

an extensive annotation approach was applied to assign functional information to

over 800,000 putative transcripts. The annotation is based on direct searches

against public databases, including RefSeq and InterPro. Gene Ontology (GO),

Enzyme Commission (EC) annotations and associated Kyoto Encyclopedia of Genes and

Genomes (KEGG) pathway maps are also collected. As a proof-of-concept, the

selection of biosynthetic gene candidates associated with six specialized

metabolic pathways is described. A web-based BLAST server has been established to

allow public access to assembled transcriptome databases for all 75 plant species

of the PhytoMetaSyn Project (www.phytometasyn.ca).

Copyright © 2013 The Authors. Published by Elsevier B.V. All rights reserved.

DOI: 10.1016/j.jbiotec.2013.04.004

PMID: 23602801 [Indexed for MEDLINE]

1057. PLoS One. 2013 Jul 3;8(7):e67434. doi: 10.1371/journal.pone.0067434. Print 2013.

LegumeGRN: a gene regulatory network prediction server for functional and

comparative studies.

Wang M(1), Verdier J, Benedito VA, Tang Y, Murray JD, Ge Y, Becker JD, Carvalho

H, Rogers C, Udvardi M, He J.

Author information:

(1)Division of Plant Biology, The Samuel Roberts Noble Foundation, Ardmore,

Oklahoma, United States of America. mwang@noble.org

Building accurate gene regulatory networks (GRNs) from high-throughput gene

expression data is a long-standing challenge. However, with the emergence of new

algorithms combined with the increase of transcriptomic data availability, it is

now reachable. To help biologists to investigate gene regulatory relationships,

we developed a web-based computational service to build, analyze and visualize

GRNs that govern various biological processes. The web server is preloaded with

all available Affymetrix GeneChip-based transcriptomic and annotation data from

the three model legume species, i.e., Medicago truncatula, Lotus japonicus and

Glycine max. Users can also upload their own transcriptomic and transcription

factor datasets from any other species/organisms to analyze their in-house

experiments. Users are able to select which experiments, genes and algorithms

they will consider to perform their GRN analysis. To achieve this flexibility and

improve prediction performance, we have implemented multiple mainstream GRN

prediction algorithms including co-expression, Graphical Gaussian Models (GGMs),

Context Likelihood of Relatedness (CLR), and parallelized versions of TIGRESS and

GENIE3. Besides these existing algorithms, we also proposed a parallel Bayesian

network learning algorithm, which can infer causal relationships (i.e.,

directionality of interaction) and scale up to several thousands of genes.

Moreover, this web server also provides tools to allow integrative and

comparative analysis between predicted GRNs obtained from different algorithms or

experiments, as well as comparisons between legume species. The web site is

available at http://legumegrn.noble.org.

DOI: 10.1371/journal.pone.0067434

PMCID: PMC3701055

PMID: 23844010 [Indexed for MEDLINE]

1058. Bioinformatics. 2013 Jul 1;29(13):1654-62. doi: 10.1093/bioinformatics/btt202.

Epub 2013 May 21.

NETAL: a new graph-based method for global alignment of protein-protein

interaction networks.

Neyshabur B(1), Khadem A, Hashemifar S, Arab SS.

Author information:

(1)Department of Computer Engineering, Sharif University of Technology, Tehran,

Iran.

MOTIVATION: The interactions among proteins and the resulting networks of such

interactions have a central role in cell biology. Aligning these networks gives

us important information, such as conserved complexes and evolutionary

relationships. Although there have been several publications on the global

alignment of protein networks; however, none of proposed methods are able to

produce a highly conserved and meaningful alignment. Moreover, time complexity of

current algorithms makes them impossible to use for multiple alignment of several

large networks together.

RESULTS: We present a novel algorithm for the global alignment of protein-protein

interaction networks. It uses a greedy method, based on the alignment scoring

matrix, which is derived from both biological and topological information of

input networks to find the best global network alignment. NETAL outperforms other

global alignment methods in terms of several measurements, such as Edge

Correctness, Largest Common Connected Subgraphs and the number of common Gene

Ontology terms between aligned proteins. As the running time of NETAL is much

less than other available methods, NETAL can be easily expanded to multiple

alignment algorithm. Furthermore, NETAL overpowers all other existing algorithms

in term of performance so that the short running time of NETAL allowed us to

implement it as the first server for global alignment of protein-protein

interaction networks.

AVAILABILITY: Binaries supported on linux are freely available for download at

http://www.bioinf.cs.ipm.ir/software/netal.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt202

PMID: 23696650 [Indexed for MEDLINE]

1059. Bioinformatics. 2013 Jul 1;29(13):1698-9. doi: 10.1093/bioinformatics/btt262.

Epub 2013 May 9.

pyDockWEB: a web server for rigid-body protein-protein docking using

electrostatics and desolvation scoring.

Jiménez-García B(1), Pons C, Fernández-Recio J.

Author information:

(1)Joint BSC-IRB Research Programme in Computational Biology, Department of Life

Sciences, Barcelona Supercomputing Center, National Institute of Bioinformatics,

Jordi Girona 29, 08034 Barcelona, Spain.

pyDockWEB is a web server for the rigid-body docking prediction of

protein-protein complex structures using a new version of the pyDock scoring

algorithm. We use here a new custom parallel FTDock implementation, with adjusted

grid size for optimal FFT calculations, and a new version of pyDock, which

dramatically speeds up calculations while keeping the same predictive accuracy.

Given the 3D coordinates of two interacting proteins, pyDockWEB returns the best

docking orientations as scored mainly by electrostatics and desolvation

energy.AVAILABILITY AND IMPLEMENTATION: The server does not require registration

by the user and is freely accessible for academics at

http://life.bsc.es/servlet/pydock.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt262

PMID: 23661696 [Indexed for MEDLINE]

1060. Bioinformatics. 2013 Jul 1;29(13):1693-5. doi: 10.1093/bioinformatics/btt265.

Epub 2013 May 8.

DAPPLE: a pipeline for the homology-based prediction of phosphorylation sites.

Trost B(1), Arsenault R, Griebel P, Napper S, Kusalik A.

Author information:

(1)Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N

5C9, Canada. brett.trost@usask.ca

SUMMARY: While many experimentally characterized phosphorylation sites exist for

certain organisms, such as human, rat and mouse, few sites are known for other

organisms, hampering related research efforts. We have developed a software

pipeline called DAPPLE that automates the process of using known phosphorylation

sites from other organisms to identify putative sites in an organism of interest.

AVAILABILITY: DAPPLE is available as a web server at http://saphire.usask.ca.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt265

PMID: 23658419 [Indexed for MEDLINE]

1061. Bioinformatics. 2013 Jul 1;29(13):1614-22. doi: 10.1093/bioinformatics/btt196.

Epub 2013 Apr 26.

Incorporating key position and amino acid residue features to identify general

and species-specific Ubiquitin conjugation sites.

Chen X(1), Qiu JD, Shi SP, Suo SB, Huang SY, Liang RP.

Author information:

(1)Department of Chemistry, Nanchang University, Nanchang 330031, People's

Republic of China.

MOTIVATION: Systematic dissection of the ubiquitylation proteome is emerging as

an appealing but challenging research topic because of the significant roles

ubiquitylation play not only in protein degradation but also in many other

cellular functions. High-throughput experimental studies using mass spectrometry

have identified many ubiquitylation sites, primarily from eukaryotes. However,

the vast majority of ubiquitylation sites remain undiscovered, even in

well-studied systems. Because mass spectrometry-based experimental approaches for

identifying ubiquitylation events are costly, time-consuming and biased toward

abundant proteins and proteotypic peptides, in silico prediction of

ubiquitylation sites is a potentially useful alternative strategy for whole

proteome annotation. Because of various limitations, current ubiquitylation site

prediction tools were not well designed to comprehensively assess proteomes.

RESULTS: We present a novel tool known as UbiProber, specifically designed for

large-scale predictions of both general and species-specific ubiquitylation

sites. We collected proteomics data for ubiquitylation from multiple species from

several reliable sources and used them to train prediction models by a

comprehensive machine-learning approach that integrates the information from key

positions and key amino acid residues. Cross-validation tests reveal that

UbiProber achieves some improvement over existing tools in predicting

species-specific ubiquitylation sites. Moreover, independent tests show that

UbiProber improves the areas under receiver operating characteristic curves by

~15% by using the Combined model.

AVAILABILITY: The UbiProber server is freely available on the web at

http://bioinfo.ncu.edu.cn/UbiProber.aspx. The software system of UbiProber can be

downloaded at the same site.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt196

PMID: 23626001 [Indexed for MEDLINE]

1062. BMC Bioinformatics. 2013 Jul 1;14:211. doi: 10.1186/1471-2105-14-211.

Jenner-predict server: prediction of protein vaccine candidates (PVCs) in

bacteria based on host-pathogen interactions.

Jaiswal V(1), Chanumolu SK, Gupta A, Chauhan RS, Rout C.

Author information:

(1)Department of Biotechnology and Bioinformatics, Jaypee University of

Information Technology, Waknaghat, Solan, Himachal Pradesh 173234, India.

BACKGROUND: Subunit vaccines based on recombinant proteins have been effective in

preventing infectious diseases and are expected to meet the demands of future

vaccine development. Computational approach, especially reverse vaccinology (RV)

method has enormous potential for identification of protein vaccine candidates

(PVCs) from a proteome. The existing protective antigen prediction software and

web servers have low prediction accuracy leading to limited applications for

vaccine development. Besides machine learning techniques, those software and web

servers have considered only protein's adhesin-likeliness as criterion for

identification of PVCs. Several non-adhesin functional classes of proteins

involved in host-pathogen interactions and pathogenesis are known to provide

protection against bacterial infections. Therefore, knowledge of bacterial

pathogenesis has potential to identify PVCs.

RESULTS: A web server, Jenner-Predict, has been developed for prediction of PVCs

from proteomes of bacterial pathogens. The web server targets host-pathogen

interactions and pathogenesis by considering known functional domains from

protein classes such as adhesin, virulence, invasin, porin, flagellin,

colonization, toxin, choline-binding, penicillin-binding, transferring-binding,

fibronectin-binding and solute-binding. It predicts non-cytosolic proteins

containing above domains as PVCs. It also provides vaccine potential of PVCs in

terms of their possible immunogenicity by comparing with experimentally known

IEDB epitopes, absence of autoimmunity and conservation in different strains.

Predicted PVCs are prioritized so that only few prospective PVCs could be

validated experimentally. The performance of web server was evaluated against

known protective antigens from diverse classes of bacteria reported in Protegen

database and datasets used for VaxiJen server development. The web server

efficiently predicted known vaccine candidates reported from Streptococcus

pneumoniae and Escherichia coli proteomes. The Jenner-Predict server outperformed

NERVE, Vaxign and VaxiJen methods. It has sensitivity of 0.774 and 0.711 for

Protegen and VaxiJen dataset, respectively while specificity of 0.940 has been

obtained for the latter dataset.

CONCLUSIONS: Better prediction accuracy of Jenner-Predict web server signifies

that domains involved in host-pathogen interactions and pathogenesis are better

criteria for prediction of PVCs. The web server has successfully predicted

maximum known PVCs belonging to different functional classes. Jenner-Predict

server is freely accessible at http://117.211.115.67/vaccine/home.html.

DOI: 10.1186/1471-2105-14-211

PMCID: PMC3701604

PMID: 23815072 [Indexed for MEDLINE]

1063. Curr Protoc Mol Biol. 2013 Jul;Chapter 12:Unit 12.16. doi:

10.1002/0471142727.mb1216s103.

Engineering customized TALE nucleases (TALENs) and TALE transcription factors by

fast ligation-based automatable solid-phase high-throughput (FLASH) assembly.

Reyon D(1), Maeder ML, Khayter C, Tsai SQ, Foley JE, Sander JD, Joung JK.

Author information:

(1)Molecular Pathology Unit, Center for Computational and Integrative Biology,

Massachusetts General Hospital, Charlestown, Massachusetts, USA.

Customized DNA-binding domains made using transcription activator-like effector

(TALE) repeats are rapidly growing in importance as widely applicable research

tools. TALE nucleases (TALENs), composed of an engineered array of TALE repeats

fused to the FokI nuclease domain, have been used successfully for directed

genome editing in various organisms and cell types. TALE transcription factors

(TALE-TFs), consisting of engineered TALE repeat arrays linked to a

transcriptional regulatory domain, have been used to up- or downregulate

expression of endogenous genes in human cells and plants. This unit describes a

detailed protocol for the recently described fast ligation-based automatable

solid-phase high-throughput (FLASH) assembly method. FLASH enables automated

high-throughput construction of engineered TALE repeats using an automated liquid

handling robot or manually using a multichannel pipet. Using the automated

approach, a single researcher can construct up to 96 DNA fragments encoding TALE

repeat arrays of various lengths in a single day, and then clone these to

construct sequence-verified TALEN or TALE-TF expression plasmids in a week or

less. Plasmids required for FLASH are available by request from the Joung lab

(http://eGenome.org). This unit also describes improvements to the Zinc Finger

and TALE Targeter (ZiFiT Targeter) web server (http://ZiFiT.partners.org) that

facilitate the design and construction of FLASH TALE repeat arrays in high

throughput.

© 2013 by John Wiley & Sons, Inc.

DOI: 10.1002/0471142727.mb1216s103

PMCID: PMC3767754

PMID: 23821439 [Indexed for MEDLINE]

1064. IEEE/ACM Trans Comput Biol Bioinform. 2013 Jul-Aug;10(4):994-1008. doi:

10.1109/TCBB.2013.104.

Designing template-free predictor for targeting protein-ligand binding sites with

classifier ensemble and spatial clustering.

Yu DJ(1), Hu J(1), Yang J(2), Shen HB(2), Tang J(1), Yang JY(1).

Author information:

(1)Nanjing University of Science and Technology, Nanjing. (2)Shanghai Jiao Tong

University, Shanghai and Ministry of Education of China, Shanghai.

Accurately identifying the protein-ligand binding sites or pockets is of

significant importance for both protein function analysis and drug design.

Although much progress has been made, challenges remain, especially when the 3D

structures of target proteins are not available or no homology templates can be

found in the library, where the template-based methods are hard to be applied. In

this paper, we report a new ligand-specific template-free predictor called

TargetS for targeting protein-ligand binding sites from primary sequences.

TargetS first predicts the binding residues along the sequence with

ligand-specific strategy and then further identifies the binding sites from the

predicted binding residues through a recursive spatial clustering algorithm.

Protein evolutionary information, predicted protein secondary structure, and

ligand-specific binding propensities of residues are combined to construct

discriminative features; an improved AdaBoost classifier ensemble scheme based on

random undersampling is proposed to deal with the serious imbalance problem

between positive (binding) and negative (nonbinding) samples. Experimental

results demonstrate that TargetS achieves high performances and outperforms many

existing predictors. TargetS web server and data sets are freely available at:

http://www.csbio.sjtu.edu.cn/bioinf/TargetS/ for academic use.

DOI: 10.1109/TCBB.2013.104

PMID: 24334392 [Indexed for MEDLINE]

1065. J Neurol. 2013 Jul;260(7):1770-7. doi: 10.1007/s00415-013-6872-8. Epub 2013 Mar

3.

An algorithm for candidate sequencing in non-dystrophic skeletal muscle

channelopathies.

Nam TS(1), Lossin C, Kim DU, Kim MK, Kim YO, Choi KH, Choi SY, Park SC, Na IS.

Author information:

(1)Department of Neurology, The Brain Korea 21 Project, Chonnam National

University Medical School, 8 Hakdong, Donggu, Gwangju 501-757, Korea.

Human skeletal muscle channelopathies (HSMCs) are a group of heritable conditions

with ion channel-related etiology and similar presentation. To create a

comprehensive picture of the phenotypic spectrum for each condition and to devise

a strategy that facilitates the differential diagnosis, we collected the genotype

and phenotype information from more than 500 previously published HSMC studies.

Using these records, we were able to identify clear correlations between

particular clinical features and the underlying alteration(s) in the genes SCN4A,

CACNA1S, KCNJ2, and CLCN1. This allowed us to develop a clinical, symptom-based,

binary decision flow algorithm that predicts the proper genetic origin with high

accuracy (0.88-0.93). The algorithm was implemented in a stand-alone online tool

("CGPS"- http://cgps.ddd.co.kr ) to assist with HSCM diagnosis in the clinical

practice. The CGPS provides simple, symptom-oriented navigation that guides the

user to the most likely molecular basis of the presentation, which permits highly

targeted genetic screens and, upon confirmation, tailored pharmacotherapy based

on the molecular origin.

DOI: 10.1007/s00415-013-6872-8

PMID: 23456025 [Indexed for MEDLINE]

1066. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W115-22. doi:

10.1093/nar/gkt533.

GeneMANIA prediction server 2013 update.

Zuberi K(1), Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, Morris Q.

Author information:

(1)The Donnelly Centre, University of Toronto, Ontario, Canada.

GeneMANIA (http://www.genemania.org) is a flexible user-friendly web interface

for generating hypotheses about gene function, analyzing gene lists and

prioritizing genes for functional assays. Given a query gene list, GeneMANIA

extends the list with functionally similar genes that it identifies using

available genomics and proteomics data. GeneMANIA also reports weights that

indicate the predictive value of each selected data set for the query. GeneMANIA

can also be used in a function prediction setting: given a query gene, GeneMANIA

finds a small set of genes that are most likely to share function with that gene

based on their interactions with it. Enriched Gene Ontology categories among this

set can sometimes point to the function of the gene. Seven organisms are

currently supported (Arabidopsis thaliana, Caenorhabditis elegans, Drosophila

melanogaster, Mus musculus, Homo sapiens, Rattus norvegicus and Saccharomyces

cerevisiae), and hundreds of data sets have been collected from GEO, BioGRID,

IRefIndex and I2D, as well as organism-specific functional genomics data sets.

Users can customize their search by selecting specific data sets to query and by

uploading their own data sets to analyze.

DOI: 10.1093/nar/gkt533

PMCID: PMC3692113

PMID: 23794635 [Indexed for MEDLINE]

1067. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W544-56. doi:

10.1093/nar/gkt519. Epub 2013 Jun 14.

kmer-SVM: a web server for identifying predictive regulatory sequence features in

genomic data sets.

Fletez-Brant C(1), Lee D, McCallion AS, Beer MA.

Author information:

(1)McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University

School of Medicine, Baltimore, MD 21205, USA.

Massively parallel sequencing technologies have made the generation of genomic

data sets a routine component of many biological investigations. For example,

Chromatin immunoprecipitation followed by sequence assays detect genomic regions

bound (directly or indirectly) by specific factors, and DNase-seq identifies

regions of open chromatin. A major bottleneck in the interpretation of these data

is the identification of the underlying DNA sequence code that defines, and

ultimately facilitates prediction of, these transcription factor (TF) bound or

open chromatin regions. We have recently developed a novel computational

methodology, which uses a support vector machine (SVM) with kmer sequence

features (kmer-SVM) to identify predictive combinations of short transcription

factor-binding sites, which determine the tissue specificity of these genomic

assays (Lee, Karchin and Beer, Discriminative prediction of mammalian enhancers

from DNA sequence. Genome Res. 2011; 21:2167-80). This regulatory information can

(i) give confidence in genomic experiments by recovering previously known binding

sites, and (ii) reveal novel sequence features for subsequent experimental

testing of cooperative mechanisms. Here, we describe the development and

implementation of a web server to allow the broader research community to

independently apply our kmer-SVM to analyze and interpret their genomic datasets.

We analyze five recently published data sets and demonstrate how this tool

identifies accessory factors and repressive sequence elements. kmer-SVM is

available at http://kmersvm.beerlab.org.

DOI: 10.1093/nar/gkt519

PMCID: PMC3692045

PMID: 23771147 [Indexed for MEDLINE]

1068. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W314-21. doi:

10.1093/nar/gkt503. Epub 2013 Jun 12.

Depth: a web server to compute depth, cavity sizes, detect potential

small-molecule ligand-binding cavities and predict the pKa of ionizable residues

in proteins.

Tan KP(1), Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS.

Author information:

(1)Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore

138671,

Residue depth accurately measures burial and parameterizes local protein

environment. Depth is the distance of any atom/residue to the closest bulk water.

We consider the non-bulk waters to occupy cavities, whose volumes are determined

using a Voronoi procedure. Our estimation of cavity sizes is statistically

superior to estimates made by CASTp and VOIDOO, and on par with McVol over a data

set of 40 cavities. Our calculated cavity volumes correlated best with the

experimentally determined destabilization of 34 mutants from five proteins. Some

of the cavities identified are capable of binding small molecule ligands. In this

study, we have enhanced our depth-based predictions of binding sites by including

evolutionary information. We have demonstrated that on a database (LigASite) of

∼200 proteins, we perform on par with ConCavity and better than MetaPocket 2.0.

Our predictions, while less sensitive, are more specific and precise. Finally, we

use depth (and other features) to predict pKas of GLU, ASP, LYS and HIS residues.

Our results produce an average error of just <1 pH unit over 60 predictions. Our

simple empirical method is statistically on par with two and superior to three

other methods while inferior to only one. The DEPTH server

(http://mspc.bii.a-star.edu.sg/depth/) is an ideal tool for rapid yet accurate

structural analyses of protein structures.

DOI: 10.1093/nar/gkt503

PMCID: PMC3692129

PMID: 23766289 [Indexed for MEDLINE]

1069. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W486-91. doi:

10.1093/nar/gkt486. Epub 2013 Jun 12.

RNAtips: Analysis of temperature-induced changes of RNA secondary structure.

Chursov A(1), Kopetzky SJ, Bocharov G, Frishman D, Shneider A.

Author information:

(1)Department of Genome Oriented Bioinformatics, Technische Universität München,

Wissenschaftzentrum Weihenstephan, Maximus-von-Imhof-Forum 3, D-85354 Freising,

Germany.

Although multiple biological phenomena are related to temperature (e.g. elevation

of body temperature due to an illness, adaptation to environmental temperature

conditions, biology of coldblooded versus warm-blooded organisms), the molecular

mechanisms of these processes remain to be understood. Perturbations of secondary

RNA structures may play an important role in an organism's reaction to

temperature change--in all organisms from viruses and bacteria to humans. Here,

we present RNAtips (temperature-induced perturbation of structure) web server,

which can be used to predict regions of RNA secondary structures that are likely

to undergo structural alterations prompted by temperature change. The server can

also be used to: (i) detect those regions in two homologous RNA sequences that

undergo different structural perturbations due to temperature change and (ii)

test whether these differences are specific to the particular nucleotide

substitutions distinguishing the sequences. The RNAtips web server is freely

accessible without any login requirement at http://rnatips.org.

DOI: 10.1093/nar/gkt486

PMCID: PMC3692058

PMID: 23766288 [Indexed for MEDLINE]

1070. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W303-7. doi: 10.1093/nar/gkt498.

Epub 2013 Jun 12.

The FunFOLD2 server for the prediction of protein-ligand interactions.

Roche DB(1), Buenavista MT, McGuffin LJ.

Author information:

(1)Laboratoire de génomique et biochimie du métabolisme, Genoscope, Institut de

Génomique, Commissariat à l'Energie Atomique et aux Energies Alternatives, Evry,

Essonne 91057, France. danroche@genoscope.cns.fr

The FunFOLD2 server is a new independent server that integrates our novel

protein-ligand binding site and quality assessment protocols for the prediction

of protein function (FN) from sequence via structure. Our guiding principles

were, first, to provide a simple unified resource to make our function prediction

software easily accessible to all via a simple web interface and, second, to

produce integrated output for predictions that can be easily interpreted. The

server provides a clean web interface so that results can be viewed on a single

page and interpreted by non-experts at a glance. The output for the prediction is

an image of the top predicted tertiary structure annotated to indicate putative

ligand-binding site residues. The results page also includes a list of the most

likely binding site residues and the types of predicted ligands and their

frequencies in similar structures. The protein-ligand interactions can also be

interactively visualized in 3D using the Jmol plug-in. The raw machine readable

data are provided for developers, which comply with the Critical Assessment of

Techniques for Protein Structure Prediction data standards for FN predictions.

The FunFOLD2 webserver is freely available to all at the following web site:

http://www.reading.ac.uk/bioinf/FunFOLD/FunFOLD\_form\_2\_0.html.

DOI: 10.1093/nar/gkt498

PMCID: PMC3692132

PMID: 23761453 [Indexed for MEDLINE]

1071. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W142-9. doi: 10.1093/nar/gkt496.

Epub 2013 Jun 12.

PGMRA: a web server for (phenotype x genotype) many-to-many relation analysis in

GWAS.

Arnedo J(1), del Val C, de Erausquin GA, Romero-Zaliz R, Svrakic D, Cloninger CR,

Zwir I.

Author information:

(1)Department of Computer Science and Artificial Intelligence, University of

Granada, E-18071 Granada, Spain.

It has been proposed that single nucleotide polymorphisms (SNPs) discovered by

genome-wide association studies (GWAS) account for only a small fraction of the

genetic variation of complex traits in human population. The remaining

unexplained variance or missing heritability is thought to be due to marginal

effects of many loci with small effects and has eluded attempts to identify its

sources. Combination of different studies appears to resolve in part this

problem. However, neither individual GWAS nor meta-analytic combinations thereof

are helpful for disclosing which genetic variants contribute to explain a

particular phenotype. Here, we propose that most of the missing heritability is

latent in the GWAS data, which conceals intermediate phenotypes. To uncover such

latent information, we propose the PGMRA server that introduces phenomics--the

full set of phenotype features of an individual--to identify SNP-set structures

in a broader sense, i.e. causally cohesive genotype-phenotype relations. These

relations are agnostically identified (without considering disease status of the

subjects) and organized in an interpretable fashion. Then, by incorporating a

posteriori the subject status within each relation, we can establish the risk

surface of a disease in an unbiased mode. This approach complements-instead of

replaces-current analysis methods. The server is publically available at

http://phop.ugr.es/fenogeno.

DOI: 10.1093/nar/gkt496

PMCID: PMC3692099

PMID: 23761451 [Indexed for MEDLINE]

1072. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W373-8. doi: 10.1093/nar/gkt509.

Epub 2013 Jun 12.

VLDP web server: a powerful geometric tool for analysing protein structures in

their environment.

Esque J(1), Léonard S, de Brevern AG, Oguey C.

Author information:

(1)LPTM, CNRS UMR 8089, Université Cergy-Pontoise, F-95302 Cergy-Pontoise,

France.

Protein structures are an ensemble of atoms determined experimentally mostly by

X-ray crystallography or Nuclear Magnetic Resonance. Studying 3D protein

structures is a key point for better understanding protein function at a

molecular level. We propose a set of accurate tools, for analysing protein

structures, based on the reliable method of Voronoi-Laguerre tessellations. The

Voronoi Laguerre Delaunay Protein web server (VLDPws) computes the Laguerre

tessellation on a whole given system first embedded in solvent. Through this fine

description, VLDPws gives the following data: (i) Amino acid volumes evaluated

with high precision, as confirmed by good correlations with experimental data.

(ii) A novel definition of inter-residue contacts within the given protein. (iii)

A measure of the residue exposure to solvent that significantly improves the

standard notion of accessibility in some cases. At present, no equivalent web

server is available. VLDPws provides output in two complementary forms: direct

visualization of the Laguerre tessellation, mostly its polygonal molecular

surfaces; files of volumes; and areas, contacts and similar data for each residue

and each atom. These files are available for download for further analysis.

VLDPws can be accessed at http://www.dsimb.inserm.fr/dsimb\_tools/vldp.

DOI: 10.1093/nar/gkt509

PMCID: PMC3692094

PMID: 23761450 [Indexed for MEDLINE]

1073. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W492-8. doi: 10.1093/nar/gkt501.

Epub 2013 Jun 12.

HiTRACE-Web: an online tool for robust analysis of high-throughput capillary

electrophoresis.

Kim H(1), Cordero P, Das R, Yoon S.

Author information:

(1)Department of Electrical and Computer Engineering, Seoul National University,

Seoul 151-744, Korea.

To facilitate the analysis of large-scale high-throughput capillary

electrophoresis data, we previously proposed a suite of efficient analysis

software named HiTRACE (High Throughput Robust Analysis of Capillary

Electrophoresis). HiTRACE has been used extensively for quantitating data from

RNA and DNA structure mapping experiments, including mutate-and-map contact

inference, chromatin footprinting, the Eterna RNA design project and other

high-throughput applications. However, HiTRACE is based on a suite of

command-line MATLAB scripts that requires nontrivial efforts to learn, use and

extend. Here, we present HiTRACE-Web, an online version of HiTRACE that includes

standard features previously available in the command-line version and additional

features such as automated band annotation and flexible adjustment of

annotations, all via a user-friendly environment. By making use of

parallelization, the on-line workflow is also faster than software

implementations available to most users on their local computers. Free access:

http://hitrace.org.

DOI: 10.1093/nar/gkt501

PMCID: PMC3692083

PMID: 23761448 [Indexed for MEDLINE]

1074. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W198-203. doi:

10.1093/nar/gkt532. Epub 2013 Jun 12.

ResponseNet2.0: Revealing signaling and regulatory pathways connecting your

proteins and genes--now with human data.

Basha O(1), Tirman S, Eluk A, Yeger-Lotem E.

Author information:

(1)Department of Clinical Biochemistry & Pharmacology, Ben-Gurion University of

the Negev, Beer-Sheva 84105, Israel.

Genome sequencing and transcriptomic profiling are two widely used approaches for

the identification of human disease pathways. However, each approach typically

provides a limited view of disease pathways: Genome sequencing can identify

disease-related mutations but rarely reveals their mode-of-action, while

transcriptomic assays do not reveal the series of events that lead to the

transcriptomic change. ResponseNet is an integrative network-optimization

approach that we developed to fill these gaps by highlighting major signaling and

regulatory molecular interaction paths that connect disease-related mutations and

genes. The ResponseNet web-server provides a user-friendly interface to

ResponseNet. Specifically, users can upload weighted lists of proteins and genes

and obtain a sparse, weighted, molecular interaction subnetwork connecting them,

that is biased toward regulatory and signaling pathways. ResponseNet2.0 enhances

the functionality of the ResponseNet web-server in two important ways. First, it

supports analysis of human data by offering a human interactome composed of

proteins, genes and micro-RNAs. Second, it offers a new informative view of the

output, including a randomization analysis, to help users assess the biological

relevance of the output subnetwork. ResponseNet2.0 is available at

http://netbio.bgu.ac.il/respnet .

DOI: 10.1093/nar/gkt532

PMCID: PMC3692079

PMID: 23761447 [Indexed for MEDLINE]

1075. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W180-6. doi: 10.1093/nar/gkt463.

Epub 2013 Jun 10.

RBPmotif: a web server for the discovery of sequence and structure preferences of

RNA-binding proteins.

Kazan H(1), Morris Q.

Author information:

(1)Department of Computer Engineering, Antalya International University, Antalya,

07190, Turkey. hilal.kazan@antalya.edu.tr

RBPmotif web server (http://www.rnamotif.org) implements tools to identify

binding preferences of RNA-binding proteins (RBPs). Given a set of sequences that

are known to be bound and unbound by the RBP of interest, RBPmotif provides two

types of analysis: (i) de novo motif finding when there is no a priori knowledge

on RBP's binding preferences and (ii) analysis of structure preferences when

there is a previously identified sequence motif for the RBP. De novo motif

finding is performed with the previously published RNAcontext algorithm that

learns discriminative motif models to identify both sequence and structure

preferences. The results of this analysis include the inferred binding

preferences of the RBP and the added predictive value of incorporating structure

preferences. Second type of analysis investigates whether the instances of the

previously identified sequence motif are enriched in a particular structure

context in bound sequences, relative to its instances in unbound sequences. On

completion, the results page shows the comparison of structure contexts of the

motif instances between bound and unbound sequences and an assessment of

statistical significance of detected preferences. In summary, RBPmotif web server

enables the concurrent analysis of sequence and structure preferences of RBPs

through a user-friendly interface.

DOI: 10.1093/nar/gkt463

PMCID: PMC3692078

PMID: 23754853 [Indexed for MEDLINE]

1076. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W84-8. doi: 10.1093/nar/gkt516.

Epub 2013 Jun 10.

FIDEA: a server for the functional interpretation of differential expression

analysis.

D'Andrea D(1), Grassi L, Mazzapioda M, Tramontano A.

Author information:

(1)Department of Physics, Sapienza University, Rome, 00185, Italy.

The results of differential expression analyses provide scientists with hundreds

to thousands of differentially expressed genes that need to be interpreted in

light of the biology of the specific system under study. This requires mapping

the genes to functional classifications that can be, for example, the KEGG

pathways or InterPro families they belong to, their GO Molecular Function,

Biological Process or Cellular Component. A statistically significant

overrepresentation of one or more category terms in the set of differentially

expressed genes is an essential step for the interpretation of the biological

significance of the results. Ideally, the analysis should be performed by

scientists who are well acquainted with the biological problem, as they have a

wealth of knowledge about the system and can, more easily than a

bioinformatician, discover less obvious and, therefore, more interesting

relationships. To allow experimentalists to explore their data in an easy and at

the same time exhaustive fashion within a single tool and to test their

hypothesis quickly and effortlessly, we developed FIDEA. The FIDEA server is

located at http://www.biocomputing.it/fidea; it is free and open to all users,

and there is no login requirement.

DOI: 10.1093/nar/gkt516

PMCID: PMC3692084

PMID: 23754850 [Indexed for MEDLINE]

1077. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W454-8. doi: 10.1093/nar/gkt472.

Epub 2013 Jun 8.

HIV N-linked glycosylation site analyzer and its further usage in anchored

alignment.

Shaw TI(1), Zhang M.

Author information:

(1)Institute of Bioinformatics, University of Georgia, GA 30605, USA.

N-linked glycosylation is a posttranslational modification that has significantly

contributed to the rapid evolution of HIV-1. In particular, enrichment of

N-linked glycosylation sites can be found within Envelope variable loops, regions

that play an essential role in HIV pathogenesis and immunogenicity. The web

server described here, the HIV N-linked Glycosylation Site Analyzer, was

developed to facilitate study of HIV diversity by tracking gp120 N-linked

glycosylation sites. This server provides an automated platform for mapping and

comparing variable loop N-linked glycosylation sites across populations of HIV-1

sequences. Furthermore, this server allows for refinement of HIV-1 sequence

alignment by using N-linked glycosylation sites in variable loops as alignment

anchors. Availability of this web server solves one of the difficult problems in

HIV gp120 alignment and analysis imposed by the extraordinary HIV-1 diversity.

The HIV N-linked Glycosylation Site Analyzer web server is available at

http://hivtools.publichealth.uga.edu/N-Glyco/.

DOI: 10.1093/nar/gkt472

PMCID: PMC3692120

PMID: 23748959 [Indexed for MEDLINE]

1078. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W349-57. doi:

10.1093/nar/gkt381. Epub 2013 Jun 8.

Scalable web services for the PSIPRED Protein Analysis Workbench.

Buchan DW(1), Minneci F, Nugent TC, Bryson K, Jones DT.

Author information:

(1)Department of Computer Science, University College London, Gower Street,

London, WC1E 6BT. d.buchan@cs.ucl.ac.uk

Here, we present the new UCL Bioinformatics Group's PSIPRED Protein Analysis

Workbench. The Workbench unites all of our previously available analysis methods

into a single web-based framework. The new web portal provides a greatly

streamlined user interface with a number of new features to allow users to better

explore their results. We offer a number of additional services to enable

computationally scalable execution of our prediction methods; these include SOAP

and XML-RPC web server access and new HADOOP packages. All software and services

are available via the UCL Bioinformatics Group website at

http://bioinf.cs.ucl.ac.uk/.

DOI: 10.1093/nar/gkt381

PMCID: PMC3692098

PMID: 23748958 [Indexed for MEDLINE]

1079. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W218-24. doi:

10.1093/nar/gkt473. Epub 2013 Jun 8.

The UCSC Interaction Browser: multidimensional data views in pathway context.

Wong CK(1), Vaske CJ, Ng S, Sanborn JZ, Benz SC, Haussler D, Stuart JM.

Author information:

(1)Biomolecular Engineering Department, University of California, Santa Cruz,

1156 High Street, Santa Cruz, CA 95064, USA.

High-throughput data sets such as genome-wide protein-protein interactions,

protein-DNA interactions and gene expression data have been published for several

model systems, especially for human cancer samples. The University of California,

Santa Cruz (UCSC) Interaction Browser (http://sysbio.soe.ucsc.edu/nets) is an

online tool for biologists to view high-throughput data sets simultaneously for

the analysis of functional relationships between biological entities. Users can

access several public interaction networks and functional genomics data sets

through the portal as well as upload their own networks and data sets for

analysis. Users can navigate through correlative relationships for focused sets

of genes belonging to biological pathways using a standard web browser. Using a

new visual modality called the CircleMap, multiple 'omics' data sets can be

viewed simultaneously within the context of curated, predicted, directed and

undirected regulatory interactions. The Interaction Browser provides an

integrative viewing of biological networks based on the consensus of many

observations about genes and their products, which may provide new insights about

normal and disease processes not obvious from any isolated data set.

DOI: 10.1093/nar/gkt473

PMCID: PMC3692096

PMID: 23748957 [Indexed for MEDLINE]

1080. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W41-6. doi: 10.1093/nar/gkt530.

Epub 2013 Jun 8.

Genome Maps, a new generation genome browser.

Medina I(1), Salavert F, Sanchez R, de Maria A, Alonso R, Escobar P, Bleda M,

Dopazo J.

Author information:

(1)Department of Computational Genomics, Centro de Investigación Príncipe Felipe,

Valencia 46012, Spain.

Genome browsers have gained importance as more genomes and related genomic

information become available. However, the increase of information brought about

by new generation sequencing technologies is, at the same time, causing a subtle

but continuous decrease in the efficiency of conventional genome browsers. Here,

we present Genome Maps, a genome browser that implements an innovative model of

data transfer and management. The program uses highly efficient technologies from

the new HTML5 standard, such as scalable vector graphics, that optimize workloads

at both server and client sides and ensure future scalability. Thus, data

management and representation are entirely carried out by the browser, without

the need of any Java Applet, Flash or other plug-in technology installation.

Relevant biological data on genes, transcripts, exons, regulatory features,

single-nucleotide polymorphisms, karyotype and so forth, are imported from web

services and are available as tracks. In addition, several DAS servers are

already included in Genome Maps. As a novelty, this web-based genome browser

allows the local upload of huge genomic data files (e.g. VCF or BAM) that can be

dynamically visualized in real time at the client side, thus facilitating the

management of medical data affected by privacy restrictions. Finally, Genome Maps

can easily be integrated in any web application by including only a few lines of

code. Genome Maps is an open source collaborative initiative available in the

GitHub repository (https://github.com/compbio-bigdata-viz/genome-maps). Genome

Maps is available at: http://www.genomemaps.org.

DOI: 10.1093/nar/gkt530

PMCID: PMC3692043

PMID: 23748955 [Indexed for MEDLINE]

1081. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W480-5. doi: 10.1093/nar/gkt461.

Epub 2013 Jun 8.

SPARCS: a web server to analyze (un)structured regions in coding RNA sequences.

Zhang Y(1), Ponty Y, Blanchette M, Lécuyer E, Waldispühl J.

Author information:

(1)School of Computer Science & McGill Centre for Bioinformatics, McGill

University, Montréal, H3A 0C6 QC, Canada.

More than a simple carrier of the genetic information, messenger RNA (mRNA)

coding regions can also harbor functional elements that evolved to control

different post-transcriptional processes, such as mRNA splicing, localization and

translation. Functional elements in RNA molecules are often encoded by secondary

structure elements. In this aticle, we introduce Structural Profile Assignment of

RNA Coding Sequences (SPARCS), an efficient method to analyze the (secondary)

structure profile of protein-coding regions in mRNAs. First, we develop a novel

algorithm that enables us to sample uniformly the sequence landscape preserving

the dinucleotide frequency and the encoded amino acid sequence of the input mRNA.

Then, we use this algorithm to generate a set of artificial sequences that is

used to estimate the Z-score of classical structural metrics such as the sum of

base pairing probabilities and the base pairing entropy. Finally, we use these

metrics to predict structured and unstructured regions in the input mRNA

sequence. We applied our methods to study the structural profile of the ASH1

genes and recovered key structural elements. A web server implementing this

discovery pipeline is available at http://csb.cs.mcgill.ca/sparcs together with

the source code of the sampling algorithm.

DOI: 10.1093/nar/gkt461

PMCID: PMC3692110

PMID: 23748952 [Indexed for MEDLINE]

1082. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W232-7. doi: 10.1093/nar/gkt471.

Epub 2013 Jun 8.

CoPAP: Coevolution of presence-absence patterns.

Cohen O(1), Ashkenazy H, Levy Karin E, Burstein D, Pupko T.

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, Ramat Aviv 69978, Israel.

Evolutionary analysis of phyletic patterns (phylogenetic profiles) is widely used

in biology, representing presence or absence of characters such as genes,

restriction sites, introns, indels and methylation sites. The phyletic pattern

observed in extant genomes is the result of ancestral gain and loss events along

the phylogenetic tree. Here we present CoPAP (coevolution of presence-absence

patterns), a user-friendly web server, which performs accurate inference of

coevolving characters as manifested by co-occurring gains and losses. CoPAP uses

state-of-the-art probabilistic methodologies to infer coevolution and allows for

advanced network analysis and visualization. We developed a platform for

comparing different algorithms that detect coevolution, which includes simulated

data with pairs of coevolving sites and independent sites. Using these simulated

data we demonstrate that CoPAP performance is higher than alternative methods. We

exemplify CoPAP utility by analyzing coevolution among thousands of bacterial

genes across 681 genomes. Clusters of coevolving genes that were detected using

our method largely coincide with known biosynthesis pathways and cellular

modules, thus exhibiting the capability of CoPAP to infer biologically meaningful

interactions. CoPAP is freely available for use at http://copap.tau.ac.il/.

DOI: 10.1093/nar/gkt471

PMCID: PMC3692100

PMID: 23748951 [Indexed for MEDLINE]

1083. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W406-11. doi:

10.1093/nar/gkt462. Epub 2013 Jun 8.

CABS-fold: Server for the de novo and consensus-based prediction of protein

structure.

Blaszczyk M(1), Jamroz M, Kmiecik S, Kolinski A.

Author information:

(1)Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of

Warsaw, Pasteura 1, 02-093 Warsaw, Poland.

The CABS-fold web server provides tools for protein structure prediction from

sequence only (de novo modeling) and also using alternative templates (consensus

modeling). The web server is based on the CABS modeling procedures ranked in

previous Critical Assessment of techniques for protein Structure Prediction

competitions as one of the leading approaches for de novo and template-based

modeling. Except for template data, fragmentary distance restraints can also be

incorporated into the modeling process. The web server output is a coarse-grained

trajectory of generated conformations, its Jmol representation and predicted

models in all-atom resolution (together with accompanying analysis). CABS-fold

can be freely accessed at http://biocomp.chem.uw.edu.pl/CABSfold.

DOI: 10.1093/nar/gkt462

PMCID: PMC3692050

PMID: 23748950 [Indexed for MEDLINE]

1084. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W535-43. doi:

10.1093/nar/gkt448. Epub 2013 Jun 7.

PscanChIP: Finding over-represented transcription factor-binding site motifs and

their correlations in sequences from ChIP-Seq experiments.

Zambelli F(1), Pesole G, Pavesi G.

Author information:

(1)Dipartimento di Bioscienze, Università di Milano, Via Celoria 26, 20133

Milano, Italy.

Chromatin immunoprecipitation followed by sequencing with next-generation

technologies (ChIP-Seq) has become the de facto standard for building genome-wide

maps of regions bound by a given transcription factor (TF). The regions

identified, however, have to be further analyzed to determine the actual

DNA-binding sites for the TF, as well as sites for other TFs belonging to the

same TF complex or in general co-operating or interacting with it in

transcription regulation. PscanChIP is a web server that, starting from a

collection of genomic regions derived from a ChIP-Seq experiment, scans them

using motif descriptors like JASPAR or TRANSFAC position-specific frequency

matrices, or descriptors uploaded by users, and it evaluates both motif

enrichment and positional bias within the regions according to different measures

and criteria. PscanChIP can successfully identify not only the actual binding

sites for the TF investigated by a ChIP-Seq experiment but also secondary motifs

corresponding to other TFs that tend to bind the same regions, and, if present,

precise positional correlations among their respective sites. The web interface

is free for use, and there is no login requirement. It is available at

http://www.beaconlab.it/pscan\_chip\_dev.

DOI: 10.1093/nar/gkt448

PMCID: PMC3692095

PMID: 23748563 [Indexed for MEDLINE]

1085. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W187-91. doi:

10.1093/nar/gkt459. Epub 2013 Jun 5.

SPEDRE: a web server for estimating rate parameters for cell signaling dynamics

in data-rich environments.

Nim TH(1), White JK, Tucker-Kellogg L.

Author information:

(1)Computational Systems Biology Programme, Singapore-MIT Alliance, National

University of Singapore, 117576, Singapore.

Cell signaling pathways and metabolic networks are often modeled using ordinary

differential equations (ODEs) to represent the production/consumption of

molecular species over time. Regardless whether a model is built de novo or

adapted from previous models, there is a need to estimate kinetic rate constants

based on time-series experimental measurements of molecular abundance. For

data-rich cases such as proteomic measurements of all species, spline-based

parameter estimation algorithms have been developed to avoid solving all the ODEs

explicitly. We report the development of a web server for a spline-based method.

Systematic Parameter Estimation for Data-Rich Environments (SPEDRE) estimates

reaction rates for biochemical networks. As input, it takes the connectivity of

the network and the concentrations of the molecular species at discrete time

points. SPEDRE is intended for large sparse networks, such as signaling cascades

with many proteins but few reactions per protein. If data are available for all

species in the network, it provides global coverage of the parameter space, at

low resolution and with approximate accuracy. The output is an optimized value

for each reaction rate parameter, accompanied by a range and bin plot. SPEDRE

uses tools from COPASI for pre-processing and post-processing. SPEDRE is a free

service at http://LTKLab.org/SPEDRE.

DOI: 10.1093/nar/gkt459

PMCID: PMC3692124

PMID: 23742908 [Indexed for MEDLINE]

1086. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W249-55. doi:

10.1093/nar/gkt284. Epub 2013 Jun 5.

MCPath: Monte Carlo path generation approach to predict likely allosteric

pathways and functional residues.

Kaya C(1), Armutlulu A, Ekesan S, Haliloglu T.

Author information:

(1)Department of Chemical Engineering and Polymer Research Center, Bogazici

University, Bebek, 34342, Istanbul, Turkey.

Allosteric mechanism of proteins is essential in biomolecular signaling. An

important aspect underlying this mechanism is the communication pathways

connecting functional residues. Here, a Monte Carlo (MC) path generation approach

is proposed and implemented to define likely allosteric pathways through

generating an ensemble of maximum probability paths. The protein structure is

considered as a network of amino acid residues, and inter-residue interactions

are described by an atomistic potential function. PDZ domain structures are

presented as case studies. The analysis for bovine rhodopsin and three myosin

structures are also provided as supplementary case studies. The suggested

pathways and the residues constituting the pathways are maximally probable and

mostly agree with the previous studies. Overall, it is demonstrated that the

communication pathways could be multiple and intrinsically disposed, and the MC

path generation approach provides an effective tool for the prediction of key

residues that mediate the allosteric communication in an ensemble of pathways and

functionally plausible residues. The MCPath server is available at

http://safir.prc.boun.edu.tr/clbet\_server.

DOI: 10.1093/nar/gkt284

PMCID: PMC3692092

PMID: 23742907 [Indexed for MEDLINE]

1087. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W308-13. doi:

10.1093/nar/gkt457. Epub 2013 Jun 3.

webPDBinder: a server for the identification of ligand binding sites on protein

structures.

Bianchi V(1), Mangone I, Ferrè F, Helmer-Citterich M, Ausiello G.

Author information:

(1)Centre for Molecular Bioinformatics, Department of Biology, University of Rome

Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy.

The webPDBinder (http://pdbinder.bio.uniroma2.it/PDBinder) is a web server for

the identification of small ligand-binding sites in a protein structure.

webPDBinder searches a protein structure against a library of known binding sites

and a collection of control non-binding pockets. The number of similarities

identified with the residues in the two sets is then used to derive a propensity

value for each residue of the query protein associated to the likelihood that the

residue is part of a ligand binding site. The predicted binding residues can be

further refined using conservation scores derived from the multiple alignment of

the PFAM protein family. webPDBinder correctly identifies residues belonging to

the binding site in 77% of the cases and is able to identify binding pockets

starting from holo or apo structures with comparable performances. This is

important for all the real world cases where the query protein has been

crystallized without a ligand and is also difficult to obtain clear similarities

with bound pockets from holo pocket libraries. The input is either a PDB code or

a user-submitted structure. The output is a list of predicted binding pocket

residues with propensity and conservation values both in text and graphical

format.

DOI: 10.1093/nar/gkt457

PMCID: PMC3692056

PMID: 23737450 [Indexed for MEDLINE]

1088. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W384-8. doi: 10.1093/nar/gkt458.

Epub 2013 Jun 3.

GalaxyRefine: Protein structure refinement driven by side-chain repacking.

Heo L(1), Park H, Seok C.

Author information:

(1)Department of Chemistry, Seoul National University, Seoul 151-747, Korea.

The quality of model structures generated by contemporary protein structure

prediction methods strongly depends on the degree of similarity between the

target and available template structures. Therefore, the importance of improving

template-based model structures beyond the accuracy available from template

information has been emphasized in the structure prediction community. The

GalaxyRefine web server, freely available at http://galaxy.seoklab.org/refine, is

based on a refinement method that has been successfully tested in CASP10. The

method first rebuilds side chains and performs side-chain repacking and

subsequent overall structure relaxation by molecular dynamics simulation.

According to the CASP10 assessment, this method showed the best performance in

improving the local structure quality. The method can improve both global and

local structure quality on average, when used for refining the models generated

by state-of-the-art protein structure prediction servers.

DOI: 10.1093/nar/gkt458

PMCID: PMC3692086

PMID: 23737448 [Indexed for MEDLINE]

1089. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W266-72. doi:

10.1093/nar/gkt460. Epub 2013 Jun 3.

SPACER: Server for predicting allosteric communication and effects of regulation.

Goncearenco A(1), Mitternacht S, Yong T, Eisenhaber B, Eisenhaber F, Berezovsky

IN.

Author information:

(1)Computational Biology Unit and Department of Informatics, University of

Bergen, Bergen 5020, Norway.

The SPACER server provides an interactive framework for exploring allosteric

communication in proteins with different sizes, degrees of oligomerization and

function. SPACER uses recently developed theoretical concepts based on the

thermodynamic view of allostery. It proposes easily tractable and meaningful

measures that allow users to analyze the effect of ligand binding on the

intrinsic protein dynamics. The server shows potential allosteric sites and

allows users to explore communication between the regulatory and functional

sites. It is possible to explore, for instance, potential effector binding sites

in a given structure as targets for allosteric drugs. As input, the server only

requires a single structure. The server is freely available at

http://allostery.bii.a-star.edu.sg/.

DOI: 10.1093/nar/gkt460

PMCID: PMC3692057

PMID: 23737445 [Indexed for MEDLINE]

1090. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W459-64. doi:

10.1093/nar/gkt436. Epub 2013 May 31.

QARIP: a web server for quantitative proteomic analysis of regulated

intramembrane proteolysis.

Ivankov DN(1), Bogatyreva NS, Hönigschmid P, Dislich B, Hogl S, Kuhn PH, Frishman

D, Lichtenthaler SF.

Author information:

(1)Department of Genome Oriented Bioinformatics, Technische Universität München,

Wissenschaftszentrum Weihenstephan, Maximus-von-Imhof Forum 3, 85354 Freising,

Germany.

Regulated intramembrane proteolysis (RIP) is a critical mechanism for

intercellular communication and regulates the function of membrane proteins

through sequential proteolysis. RIP typically starts with ectodomain shedding of

membrane proteins by extracellular membrane-bound proteases followed by

intramembrane proteolysis of the resulting membrane-tethered fragment. However,

for the majority of RIP proteases the corresponding substrates and thus, their

functions, remain unknown. Proteome-wide identification of RIP protease

substrates is possible by mass spectrometry-based quantitative comparison of RIP

substrates or their cleavage products between different biological states.

However, this requires quantification of peptides from only the ectodomain or

cytoplasmic domain. Current analysis software does not allow matching peptides to

either domain. Here we present the QARIP (Quantitative Analysis of Regulated

Intramembrane Proteolysis) web server which matches identified peptides to the

protein transmembrane topology. QARIP allows determination of quantitative ratios

separately for the topological domains (cytoplasmic, ectodomain) of a given

protein and is thus a powerful tool for quality control, improvement of

quantitative ratios and identification of novel substrates in proteomic RIP

datasets. To our knowledge, the QARIP web server is the first tool directly

addressing the phenomenon of RIP. The web server is available at

http://webclu.bio.wzw.tum.de/qarip/. This website is free and open to all users

and there is no login requirement.

DOI: 10.1093/nar/gkt436

PMCID: PMC3692105

PMID: 23729472 [Indexed for MEDLINE]

1091. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W398-405. doi:

10.1093/nar/gkt453. Epub 2013 May 31.

BeEP Server: Using evolutionary information for quality assessment of protein

structure models.

Palopoli N(1), Lanzarotti E, Parisi G.

Author information:

(1)Departamento de Ciencia y Tecnologia, Universidad Nacional de Quilmes,

B1876BXD, Bernal, Buenos Aires, Argentina.

The BeEP Server (http://www.embnet.qb.fcen.uba.ar/embnet/beep.php) is an online

resource aimed to help in the endgame of protein structure prediction. It is able

to rank submitted structural models of a protein through an explicit use of

evolutionary information, a criterion differing from structural or energetic

considerations commonly used in other assessment programs. The idea behind BeEP

(Best Evolutionary Pattern) is to benefit from the substitution pattern derived

from structural constraints present in a set of homologous proteins adopting a

given protein conformation. The BeEP method uses a model of protein evolution

that takes into account the structure of a protein to build site-specific

substitution matrices. The suitability of these substitution matrices is assessed

through maximum likelihood calculations from which position-specific and global

scores can be derived. These scores estimate how well the structural constraints

derived from each structural model are represented in a sequence alignment of

homologous proteins. Our assessment on a subset of proteins from the Critical

Assessment of techniques for protein Structure Prediction (CASP) experiment has

shown that BeEP is capable of discriminating the models and selecting one or more

native-like structures. Moreover, BeEP is not explicitly parameterized to find

structural similarities between models and given targets, potentially helping to

explore the conformational ensemble of the native state.

DOI: 10.1093/nar/gkt453

PMCID: PMC3692104

PMID: 23729471 [Indexed for MEDLINE]

1092. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W441-7. doi: 10.1093/nar/gkt428.

Epub 2013 May 31.

PlantLoc: an accurate web server for predicting plant protein subcellular

localization by substantiality motif.

Tang S(1), Li T, Cong P, Xiong W, Wang Z, Sun J.

Author information:

(1)Department of Chemistry, Tongji University, Shanghai 200092, China.

Knowledge of subcellular localizations (SCLs) of plant proteins relates to their

functions and aids in understanding the regulation of biological processes at the

cellular level. We present PlantLoc, a highly accurate and fast webserver for

predicting the multi-label SCLs of plant proteins. The PlantLoc server has two

innovative characters: building localization motif libraries by a recursive

method without alignment and Gene Ontology information; and establishing simple

architecture for rapidly and accurately identifying plant protein SCLs without a

machine learning algorithm. PlantLoc provides predicted SCLs results, confidence

estimates and which is the substantiality motif and where it is located on the

sequence. PlantLoc achieved the highest accuracy (overall accuracy of 80.8%) of

identification of plant protein SCLs as benchmarked by using a new test dataset

compared other plant SCL prediction webservers. The ability of PlantLoc to

predict multiple sites was also significantly higher than for any other

webserver. The predicted substantiality motifs of queries also have great

potential for analysis of relationships with protein functional regions. The

PlantLoc server is available at http://cal.tongji.edu.cn/PlantLoc/.

DOI: 10.1093/nar/gkt428

PMCID: PMC3692052

PMID: 23729470 [Indexed for MEDLINE]

1093. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W322-8. doi: 10.1093/nar/gkt454.

Epub 2013 May 31.

PELE web server: atomistic study of biomolecular systems at your fingertips.

Madadkar-Sobhani A(1), Guallar V.

Author information:

(1)Joint BSC-IRB Research Program in Computational Biology, Barcelona

Supercomputing Center, Barcelona, Spain. armin.madadkar@bsc.es

PELE, Protein Energy Landscape Exploration, our novel technology based on protein

structure prediction algorithms and a Monte Carlo sampling, is capable of

modelling the all-atom protein-ligand dynamical interactions in an efficient and

fast manner, with two orders of magnitude reduced computational cost when

compared with traditional molecular dynamics techniques. PELE's heuristic

approach generates trial moves based on protein and ligand perturbations followed

by side chain sampling and global/local minimization. The collection of accepted

steps forms a stochastic trajectory. Furthermore, several processors may be run

in parallel towards a collective goal or defining several independent

trajectories; the whole procedure has been parallelized using the Message Passing

Interface. Here, we introduce the PELE web server, designed to make the whole

process of running simulations easier and more practical by minimizing input file

demand, providing user-friendly interface and producing abstract outputs (e.g.

interactive graphs and tables). The web server has been implemented in C++ using

Wt (http://www.webtoolkit.eu) and MySQL (http://www.mysql.com). The PELE web

server, accessible at http://pele.bsc.es, is free and open to all users with no

login requirement.

DOI: 10.1093/nar/gkt454

PMCID: PMC3692087

PMID: 23729469 [Indexed for MEDLINE]

1094. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W150-8. doi: 10.1093/nar/gkt456.

Epub 2013 May 30.

GWAS3D: Detecting human regulatory variants by integrative analysis of

genome-wide associations, chromosome interactions and histone modifications.

Li MJ(1), Wang LY, Xia Z, Sham PC, Wang J.

Author information:

(1)Department of Biochemistry, LKS Faculty of Medicine, The University of Hong

Kong, Hong Kong SAR, China.

Interpreting the genetic variants located in the regulatory regions, such as

enhancers and promoters, is an indispensable step to understand molecular

mechanism of complex traits. Recent studies show that genetic variants detected

by genome-wide association study (GWAS) are significantly enriched in the

regulatory regions. Therefore, detecting, annotating and prioritizing of genetic

variants affecting gene regulation are critical to our understanding of

genotype-phenotype relationships. Here, we developed a web server GWAS3D to

systematically analyze the genetic variants that could affect regulatory

elements, by integrating annotations from cell type-specific chromatin states,

epigenetic modifications, sequence motifs and cross-species conservation. The

regulatory elements are inferred from the genome-wide chromosome interaction

data, chromatin marks in 16 different cell types and 73 regulatory factors motifs

from the Encyclopedia of DNA Element project. Furthermore, we used these function

elements, as well as risk haplotype, binding affinity, conservation and P-values

reported from the original GWAS to reprioritize the genetic variants. Using

studies from low-density lipoprotein cholesterol, we demonstrated that our

reprioritizing approach was effective and cell type specific. In conclusion,

GWAS3D provides a comprehensive annotation and visualization tool to help users

interpreting their results. The web server is freely available at

http://jjwanglab.org/gwas3d.

DOI: 10.1093/nar/gkt456

PMCID: PMC3692118

PMID: 23723249 [Indexed for MEDLINE]

1095. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W333-9. doi: 10.1093/nar/gkt450.

Epub 2013 May 30.

BeAtMuSiC: Prediction of changes in protein-protein binding affinity on

mutations.

Dehouck Y(1), Kwasigroch JM, Rooman M, Gilis D.

Author information:

(1)Department of BioModelling, BioInformatics and BioProcesses, Université Libre

de Bruxelles, CP165/61, Av. Fr. Roosevelt 50, 1050 Brussels, Belgium.

ydehouck@ulb.ac.be

The ability of proteins to establish highly selective interactions with a variety

of (macro)molecular partners is a crucial prerequisite to the realization of

their biological functions. The availability of computational tools to evaluate

the impact of mutations on protein-protein binding can therefore be valuable in a

wide range of industrial and biomedical applications, and help rationalize the

consequences of non-synonymous single-nucleotide polymorphisms. BeAtMuSiC

(http://babylone.ulb.ac.be/beatmusic) is a coarse-grained predictor of the

changes in binding free energy induced by point mutations. It relies on a set of

statistical potentials derived from known protein structures, and combines the

effect of the mutation on the strength of the interactions at the interface, and

on the overall stability of the complex. The BeAtMuSiC server requires as input

the structure of the protein-protein complex, and gives the possibility to assess

rapidly all possible mutations in a protein chain or at the interface, with

predictive performances that are in line with the best current methodologies.

DOI: 10.1093/nar/gkt450

PMCID: PMC3692068

PMID: 23723246 [Indexed for MEDLINE]

1096. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W432-40. doi:

10.1093/nar/gkt431. Epub 2013 May 28.

IMAAAGINE: a webserver for searching hypothetical 3D amino acid side chain

arrangements in the Protein Data Bank.

Nadzirin N(1), Willett P, Artymiuk PJ, Firdaus-Raih M.

Author information:

(1)School of Biosciences and Biotechnology, Faculty of Science and Technology,

Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia.

We describe a server that allows the interrogation of the Protein Data Bank for

hypothetical 3D side chain patterns that are not limited to known patterns from

existing 3D structures. A minimal side chain description allows a variety of side

chain orientations to exist within the pattern, and generic side chain types such

as acid, base and hydroxyl-containing can be additionally deployed in the search

query. Moreover, only a subset of distances between the side chains need be

specified. We illustrate these capabilities in case studies involving arginine

stacks, serine-acid group arrangements and multiple catalytic triad-like

configurations. The IMAAAGINE server can be accessed at

http://mfrlab.org/grafss/imaaagine/.

DOI: 10.1093/nar/gkt431

PMCID: PMC3692123

PMID: 23716645 [Indexed for MEDLINE]

1097. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W582-6. doi: 10.1093/nar/gkt420.

Epub 2013 May 28.

BiDaS: a web-based Monte Carlo BioData Simulator based on sequence/feature

characteristics.

Paraskevopoulou MD(1), Vlachos IS, Athanasiadis E, Spyrou G.

Author information:

(1)Biomedical Informatics Unit, Biomedical Research Foundation, Academy of

Athens, 4 Soranou Ephessiou, 115 27 Athens, Greece.

BiDaS is a web-application that can generate massive Monte Carlo simulated

sequence or numerical feature data sets (e.g. dinucleotide content, composition,

transition, distribution properties) based on small user-provided data sets.

BiDaS server enables users to analyze their data and generate large amounts of:

(i) Simulated DNA/RNA and aminoacid (AA) sequences following practically

identical sequence and/or extracted feature distributions with the original data.

(ii) Simulated numerical features, presenting identical distributions, while

preserving the exact 2D or 3D between-feature correlations observed in the

original data sets. The server can project the provided sequences to

multidimensional feature spaces based on: (i) 38 DNA/RNA features describing

conformational and physicochemical nucleotide sequence features from the

B-DNA-VIDEO database, (ii) 122 DNA/RNA features based on conformational and

thermodynamic dinucleotide properties from the DiProDB database and (iii)

Pseudo-aminoacid composition of the initial sequences. To the best of our

knowledge, this is the first available web-server that allows users to generate

vast numbers of biological data sets with realistic characteristics, while

keeping between-feature associations. These data sets can be used for a wide

variety of current biological problems, such as the in-depth study of gene,

transcript, peptide and protein groups/families; the creation of large data sets

from just a few available members and the strengthening of machine learning

classifiers. All simulations use advanced Monte Carlo sampling techniques. The

BiDaS web-application is available at

http://bioserver-3.bioacademy.gr/Bioserver/BiDaS/.

DOI: 10.1093/nar/gkt420

PMCID: PMC3692108

PMID: 23716644 [Indexed for MEDLINE]

1098. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W15-21. doi: 10.1093/nar/gkt417.

Epub 2013 May 28.

R3D Align web server for global nucleotide to nucleotide alignments of RNA 3D

structures.

Rahrig RR(1), Petrov AI, Leontis NB, Zirbel CL.

Author information:

(1)Department of Mathematics and Statistics, Ohio Northern University, Ada, OH

45810, USA. r-rahrig@onu.edu

The R3D Align web server provides online access to 'RNA 3D Align' (R3D Align), a

method for producing accurate nucleotide-level structural alignments of RNA 3D

structures. The web server provides a streamlined and intuitive interface, input

data validation and output that is more extensive and easier to read and

interpret than related servers. The R3D Align web server offers a unique Gallery

of Featured Alignments, providing immediate access to pre-computed alignments of

large RNA 3D structures, including all ribosomal RNAs, as well as guidance on

effective use of the server and interpretation of the output. By accessing the

non-redundant lists of RNA 3D structures provided by the Bowling Green State

University RNA group, R3D Align connects users to structure files in the same

equivalence class and the best-modeled representative structure from each group.

The R3D Align web server is freely accessible at http://rna.bgsu.edu/r3dalign/.

DOI: 10.1093/nar/gkt417

PMCID: PMC3692076

PMID: 23716643 [Indexed for MEDLINE]

1099. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W358-62. doi:

10.1093/nar/gkt383. Epub 2013 May 28.

T-RMSD: a web server for automated fine-grained protein structural

classification.

Magis C(1), Di Tommaso P, Notredame C.

Author information:

(1)Bioinformatics and Genomics Programme, Centre For Genomic Regulation, Carrer

del Doctor Aiguader 88, 08003 Barcelona, Spain.

This article introduces the T-RMSD web server (tree-based on root-mean-square

deviation), a service allowing the online computation of structure-based protein

classification. It has been developed to address the relation between structural

and functional similarity in proteins, and it allows a fine-grained structural

clustering of a given protein family or group of structurally related proteins

using distance RMSD (dRMSD) variations. These distances are computed between all

pairs of equivalent residues, as defined by the ungapped columns within a given

multiple sequence alignment. Using these generated distance matrices (one per

equivalent position), T-RMSD produces a structural tree with support values for

each cluster node, reminiscent of bootstrap values. These values, associated with

the tree topology, allow a quantitative estimate of structural distances between

proteins or group of proteins defined by the tree topology. The clusters thus

defined have been shown to be structurally and functionally informative. The

T-RMSD web server is a free website open to all users and available at

http://tcoffee.crg.cat/apps/tcoffee/do:trmsd.

DOI: 10.1093/nar/gkt383

PMCID: PMC3692075

PMID: 23716642 [Indexed for MEDLINE]

1100. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W8-14. doi: 10.1093/nar/gkt427.

Epub 2013 May 28.

MISTIC: Mutual information server to infer coevolution.

Simonetti FL(1), Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C.

Author information:

(1)Bioinformatics Unit, Fundación Instituto Leloir, Av. Patricias Argentinas 435,

C1405BWE, Buenos Aires, Argentina.

MISTIC (mutual information server to infer coevolution) is a web server for

graphical representation of the information contained within a MSA (multiple

sequence alignment) and a complete analysis tool for Mutual Information networks

in protein families. The server outputs a graphical visualization of several

information-related quantities using a circos representation. This provides an

integrated view of the MSA in terms of (i) the mutual information (MI) between

residue pairs, (ii) sequence conservation and (iii) the residue cumulative and

proximity MI scores. Further, an interactive interface to explore and

characterize the MI network is provided. Several tools are offered for selecting

subsets of nodes from the network for visualization. Node coloring can be set to

match different attributes, such as conservation, cumulative MI, proximity MI and

secondary structure. Finally, a zip file containing all results can be

downloaded. The server is available at http://mistic.leloir.org.ar. In summary,

MISTIC allows for a comprehensive, compact, visually rich view of the information

contained within an MSA in a manner unique to any other publicly available web

server. In particular, the use of circos representation of MI networks and the

visualization of the cumulative MI and proximity MI concepts is novel.

DOI: 10.1093/nar/gkt427

PMCID: PMC3692073

PMID: 23716641 [Indexed for MEDLINE]

1101. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W523-30. doi:

10.1093/nar/gkt388. Epub 2013 May 22.

PiDNA: Predicting protein-DNA interactions with structural models.

Lin CK(1), Chen CY.

Author information:

(1)Center for Systems Biology, National Taiwan University, Taipei 106, Taiwan.

Predicting binding sites of a transcription factor in the genome is an important,

but challenging, issue in studying gene regulation. In the past decade, a large

number of protein-DNA co-crystallized structures available in the Protein Data

Bank have facilitated the understanding of interacting mechanisms between

transcription factors and their binding sites. Recent studies have shown that

both physics-based and knowledge-based potential functions can be applied to

protein-DNA complex structures to deliver position weight matrices (PWMs) that

are consistent with the experimental data. To further use the available

structural models, the proposed Web server, PiDNA, aims at first constructing

reliable PWMs by applying an atomic-level knowledge-based scoring function on

numerous in silico mutated complex structures, and then using the PWM constructed

by the structure models with small energy changes to predict the interaction

between proteins and DNA sequences. With PiDNA, the users can easily predict the

relative preference of all the DNA sequences with limited mutations from the

native sequence co-crystallized in the model in a single run. More predictions on

sequences with unlimited mutations can be realized by additional requests or file

uploading. Three types of information can be downloaded after prediction: (i) the

ranked list of mutated sequences, (ii) the PWM constructed by the favourable

mutated structures, and (iii) any mutated protein-DNA complex structure models

specified by the user. This study first shows that the constructed PWMs are

similar to the annotated PWMs collected from databases or literature. Second, the

prediction accuracy of PiDNA in detecting relatively high-specificity sites is

evaluated by comparing the ranked lists against in vitro experiments from

protein-binding microarrays. Finally, PiDNA is shown to be able to select the

experimentally validated binding sites from 10,000 random sites with high

accuracy. With PiDNA, the users can design biological experiments based on the

predicted sequence specificity and/or request mutated structure models for

further protein design. As well, it is expected that PiDNA can be incorporated

with chromatin immunoprecipitation data to refine large-scale inference of in

vivo protein-DNA interactions. PiDNA is available at:

http://dna.bime.ntu.edu.tw/pidna.

DOI: 10.1093/nar/gkt388

PMCID: PMC3692134

PMID: 23703214 [Indexed for MEDLINE]

1102. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W422-6. doi: 10.1093/nar/gkt416.

Epub 2013 May 22.

Vienna-PTM web server: a toolkit for MD simulations of protein post-translational

modifications.

Margreitter C(1), Petrov D, Zagrovic B.

Author information:

(1)Institute for Molecular Modeling and Simulation, University of Natural

Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria.

Post-translational modifications (PTMs) play a key role in numerous cellular

processes by directly affecting structure, dynamics and interaction networks of

target proteins. Despite their importance, our understanding of protein PTMs at

the atomistic level is still largely incomplete. Molecular dynamics (MD)

simulations, which provide high-resolution insight into biomolecular function and

underlying mechanisms, are in principle ideally suited to tackle this problem.

However, because of the challenges associated with the development of novel MD

parameters and a general lack of suitable computational tools for incorporating

PTMs in target protein structures, MD simulations of post-translationally

modified proteins have historically lagged significantly behind the studies of

unmodified proteins. Here, we present Vienna-PTM web server

(http://vienna-ptm.univie.ac.at), a platform for automated introduction of PTMs

of choice to protein 3D structures (PDB files) in a user-friendly visual

environment. With 256 different enzymatic and non-enzymatic PTMs available, the

server performs geometrically realistic introduction of modifications at sites of

interests, as well as subsequent energy minimization. Finally, the server makes

available force field parameters and input files needed to run MD simulations of

modified proteins within the framework of the widely used GROMOS 54A7 and 45A3

force fields and GROMACS simulation package.

DOI: 10.1093/nar/gkt416

PMCID: PMC3692090

PMID: 23703210 [Indexed for MEDLINE]

1103. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W56-62. doi: 10.1093/nar/gkt437.

Epub 2013 May 22.

DNAshape: a method for the high-throughput prediction of DNA structural features

on a genomic scale.

Zhou T(1), Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, Rohs R.

Author information:

(1)Molecular and Computational Biology Program, Department of Biological

Sciences, University of Southern California, Los Angeles, CA 90089, USA.

We present a method and web server for predicting DNA structural features in a

high-throughput (HT) manner for massive sequence data. This approach provides the

framework for the integration of DNA sequence and shape analyses in genome-wide

studies. The HT methodology uses a sliding-window approach to mine DNA structural

information obtained from Monte Carlo simulations. It requires only nucleotide

sequence as input and instantly predicts multiple structural features of DNA

(minor groove width, roll, propeller twist and helix twist). The results of

rigorous validations of the HT predictions based on DNA structures solved by

X-ray crystallography and NMR spectroscopy, hydroxyl radical cleavage data,

statistical analysis and cross-validation, and molecular dynamics simulations

provide strong confidence in this approach. The DNAshape web server is freely

available at http://rohslab.cmb.usc.edu/DNAshape/.

DOI: 10.1093/nar/gkt437

PMCID: PMC3692085

PMID: 23703209 [Indexed for MEDLINE]

1104. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W281-5. doi: 10.1093/nar/gkt390.

Epub 2013 May 22.

Nucleos: a web server for the identification of nucleotide-binding sites in

protein structures.

Parca L(1), Ferré F, Ausiello G, Helmer-Citterich M.

Author information:

(1)Department of Biology, Centre for Molecular Bioinformatics, University of Rome

Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy.

Nucleos is a web server for the identification of nucleotide-binding sites in

protein structures. Nucleos compares the structure of a query protein against a

set of known template 3D binding sites representing nucleotide modules, namely

the nucleobase, carbohydrate and phosphate. Structural features, clustering and

conservation are used to filter and score the predictions. The predicted

nucleotide modules are then joined to build whole nucleotide-binding sites, which

are ranked by their score. The server takes as input either the PDB code of the

query protein structure or a user-submitted structure in PDB format. The output

of Nucleos is composed of ranked lists of predicted nucleotide-binding sites

divided by nucleotide type (e.g. ATP-like). For each ranked prediction, Nucleos

provides detailed information about the score, the template structure and the

structural match for each nucleotide module composing the nucleotide-binding

site. The predictions on the query structure and the template-binding sites can

be viewed directly on the web through a graphical applet. In 98% of the cases,

the modules composing correct predictions belong to proteins with no homology

relationship between each other, meaning that the identification of brand-new

nucleotide-binding sites is possible using information from non-homologous

proteins. Nucleos is available at http://nucleos.bio.uniroma2.it/nucleos/.

DOI: 10.1093/nar/gkt390

PMCID: PMC3692072

PMID: 23703207 [Indexed for MEDLINE]

1105. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W104-8. doi: 10.1093/nar/gkt387.

Epub 2013 May 22.

MetaRanker 2.0: a web server for prioritization of genetic variation data.

Pers TH(1), Dworzyński P, Thomas CE, Lage K, Brunak S.

Author information:

(1)Department of Systems Biology, Center for Biological Sequence Analysis,

Technical University of Denmark, Lyngby, Denmark.

MetaRanker 2.0 is a web server for prioritization of common and rare frequency

genetic variation data. Based on heterogeneous data sets including genetic

association data, protein-protein interactions, large-scale text-mining data,

copy number variation data and gene expression experiments, MetaRanker 2.0

prioritizes the protein-coding part of the human genome to shortlist candidate

genes for targeted follow-up studies. MetaRanker 2.0 is made freely available at

www.cbs.dtu.dk/services/MetaRanker-2.0.

DOI: 10.1093/nar/gkt387

PMCID: PMC3692047

PMID: 23703204 [Indexed for MEDLINE]

1106. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W465-70. doi:

10.1093/nar/gkt280. Epub 2013 May 21.

RNAiFold: a web server for RNA inverse folding and molecular design.

Garcia-Martin JA(1), Clote P, Dotu I.

Author information:

(1)Department of Biology, Boston College, Chestnut Hill, MA 02467, USA.

Synthetic biology and nanotechnology are poised to make revolutionary

contributions to the 21st century. In this article, we describe a new web server

to support in silico RNA molecular design. Given an input target RNA secondary

structure, together with optional constraints, such as requiring GC-content to

lie within a certain range, requiring the number of strong (GC), weak (AU) and

wobble (GU) base pairs to lie in a certain range, the RNAiFold web server

determines one or more RNA sequences, whose minimum free-energy secondary

structure is the target structure. RNAiFold provides access to two servers:

RNA-CPdesign, which applies constraint programming, and RNA-LNSdesign, which

applies the large neighborhood search heuristic; hence, it is suitable for larger

input structures. Both servers can also solve the RNA inverse hybridization

problem, i.e. given a representation of the desired hybridization structure,

RNAiFold returns two sequences, whose minimum free-energy hybridization is the

input target structure. The web server is publicly accessible at

http://bioinformatics.bc.edu/clotelab/RNAiFold, which provides access to two

specialized servers: RNA-CPdesign and RNA-LNSdesign. Source code for the

underlying algorithms, implemented in COMET and supported on linux, can be

downloaded at the server website.

DOI: 10.1093/nar/gkt280

PMCID: PMC3692061

PMID: 23700314 [Indexed for MEDLINE]

1107. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W389-97. doi:

10.1093/nar/gkt408. Epub 2013 May 21.

QA-RecombineIt: a server for quality assessment and recombination of protein

models.

Pawlowski M(1), Bogdanowicz A, Bujnicki JM.

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology, Trojdena 4, Warsaw PL-02-109, Poland.

marcinp@genesilico.pl

QA-RecombineIt provides a web interface to assess the quality of protein 3D

structure models and to improve the accuracy of models by merging fragments of

multiple input models. QA-RecombineIt has been developed for protein modelers who

are working on difficult problems, have a set of different homology models and/or

de novo models (from methods such as I-TASSER or ROSETTA) and would like to

obtain one consensus model that incorporates the best parts into one structure

that is internally coherent. An advanced mode is also available, in which one can

modify the operation of the fragment recombination algorithm by manually

identifying individual fragments or entire models to recombine. Our method

produces up to 100 models that are expected to be on the average more accurate

than the starting models. Therefore, our server may be useful for

crystallographic protein structure determination, where protein models are used

for Molecular Replacement to solve the phase problem. To address the latter

possibility, a special feature was added to the QA-RecombineIt server. The

QA-RecombineIt server can be freely accessed at

http://iimcb.genesilico.pl/qarecombineit/.

DOI: 10.1093/nar/gkt408

PMCID: PMC3692112

PMID: 23700309 [Indexed for MEDLINE]

1108. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W242-8. doi: 10.1093/nar/gkt399.

Epub 2013 May 18.

The PhyloFacts FAT-CAT web server: ortholog identification and function

prediction using fast approximate tree classification.

Afrasiabi C(1), Samad B, Dineen D, Meacham C, Sjölander K.

Author information:

(1)QB3 Institute, University of California, Berkeley, Berkeley, CA 94720-1762,

USA.

The PhyloFacts 'Fast Approximate Tree Classification' (FAT-CAT) web server

provides a novel approach to ortholog identification using subtree hidden Markov

model-based placement of protein sequences to phylogenomic orthology groups in

the PhyloFacts database. Results on a data set of microbial, plant and animal

proteins demonstrate FAT-CAT's high precision at separating orthologs and

paralogs and robustness to promiscuous domains. We also present results

documenting the precision of ortholog identification based on subtree hidden

Markov model scoring. The FAT-CAT phylogenetic placement is used to derive a

functional annotation for the query, including confidence scores and drill-down

capabilities. PhyloFacts' broad taxonomic and functional coverage, with >7.3 M

proteins from across the Tree of Life, enables FAT-CAT to predict orthologs and

assign function for most sequence inputs. Four pipeline parameter presets are

provided to handle different sequence types, including partial sequences and

proteins containing promiscuous domains; users can also modify individual

parameters. PhyloFacts trees matching the query can be viewed interactively

online using the PhyloScope Javascript tree viewer and are hyperlinked to various

external databases. The FAT-CAT web server is available at

http://phylogenomics.berkeley.edu/phylofacts/fatcat/.

DOI: 10.1093/nar/gkt399

PMCID: PMC3692063

PMID: 23685612 [Indexed for MEDLINE]

1109. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W192-7. doi: 10.1093/nar/gkt419.

Epub 2013 May 17.

JiffyNet: a web-based instant protein network modeler for newly sequenced

species.

Kim E(1), Kim H, Lee I.

Author information:

(1)Department of Biotechnology, College of Life Science and Biotechnology, Yonsei

University, Seoul, 120-749, Korea.

Revolutionary DNA sequencing technology has enabled affordable genome sequencing

for numerous species. Thousands of species already have completely decoded

genomes, and tens of thousands more are in progress. Naturally, parallel

expansion of the functional parts list library is anticipated, yet genome-level

understanding of function also requires maps of functional relationships, such as

functional protein networks. Such networks have been constructed for many

sequenced species including common model organisms. Nevertheless, the majority of

species with sequenced genomes still have no protein network models available.

Moreover, biologists might want to obtain protein networks for their species of

interest on completion of the genome projects. Therefore, there is high demand

for accessible means to automatically construct genome-scale protein networks

based on sequence information from genome projects only. Here, we present a

public web server, JiffyNet, specifically designed to instantly construct

genome-scale protein networks based on associalogs (functional associations

transferred from a template network by orthology) for a query species with only

protein sequences provided. Assessment of the networks by JiffyNet demonstrated

generally high predictive ability for pathway annotations. Furthermore, JiffyNet

provides network visualization and analysis pages for wide variety of molecular

concepts to facilitate network-guided hypothesis generation. JiffyNet is freely

accessible at http://www.jiffynet.org.

DOI: 10.1093/nar/gkt419

PMCID: PMC3692116

PMID: 23685435 [Indexed for MEDLINE]

1110. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W174-9. doi: 10.1093/nar/gkt407.

Epub 2013 May 17.

DRIMust: a web server for discovering rank imbalanced motifs using suffix trees.

Leibovich L(1), Paz I, Yakhini Z, Mandel-Gutfreund Y.

Author information:

(1)Department of Computer Science, Technion - Israel Institute of Technology,

Technion City, Haifa 32000, Israel.

Cellular regulation mechanisms that involve proteins and other active molecules

interacting with specific targets often involve the recognition of sequence

patterns. Short sequence elements on DNA, RNA and proteins play a central role in

mediating such molecular recognition events. Studies that focus on measuring and

investigating sequence-based recognition processes make use of statistical and

computational tools that support the identification and understanding of sequence

motifs. We present a new web application, named DRIMust, freely accessible

through the website http://drimust.technion.ac.il for de novo motif discovery

services. The DRIMust algorithm is based on the minimum hypergeometric

statistical framework and uses suffix trees for an efficient enumeration of motif

candidates. DRIMust takes as input ranked lists of sequences in FASTA format and

returns motifs that are over-represented at the top of the list, where the

determination of the threshold that defines top is data driven. The resulting

motifs are presented individually with an accurate P-value indication and as a

Position Specific Scoring Matrix. Comparing DRIMust with other state-of-the-art

tools demonstrated significant advantage to DRIMust, both in result accuracy and

in short running times. Overall, DRIMust is unique in combining efficient search

on large ranked lists with rigorous P-value assessment for the detected motifs.

DOI: 10.1093/nar/gkt407

PMCID: PMC3692051

PMID: 23685432 [Indexed for MEDLINE]

1111. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W256-65. doi:

10.1093/nar/gkt403. Epub 2013 May 16.

Catalytic site identification--a web server to identify catalytic site structural

matches throughout PDB.

Kirshner DA(1), Nilmeier JP, Lightstone FC.

Author information:

(1)Biosciences and Biotechnology Division, Physical and Life Sciences

Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA.

The catalytic site identification web server provides the innovative capability

to find structural matches to a user-specified catalytic site among all Protein

Data Bank proteins rapidly (in less than a minute). The server also can examine a

user-specified protein structure or model to identify structural matches to a

library of catalytic sites. Finally, the server provides a database of

pre-calculated matches between all Protein Data Bank proteins and the library of

catalytic sites. The database has been used to derive a set of hypothesized novel

enzymatic function annotations. In all cases, matches and putative binding sites

(protein structure and surfaces) can be visualized interactively online. The

website can be accessed at http://catsid.llnl.gov.

DOI: 10.1093/nar/gkt403

PMCID: PMC3692059

PMID: 23680785 [Indexed for MEDLINE]

1112. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W169-73. doi:

10.1093/nar/gkt393. Epub 2013 May 16.

DIANA-microT web server v5.0: service integration into miRNA functional analysis

workflows.

Paraskevopoulou MD(1), Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T,

Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG.

Author information:

(1)DIANA-Lab, Biomedical Sciences Research Center Alexander Fleming, 16672 Vari,

Greece. paraskevopoulou@fleming.gr

MicroRNAs (miRNAs) are small endogenous RNA molecules that regulate gene

expression through mRNA degradation and/or translation repression, affecting many

biological processes. DIANA-microT web server (http://www.microrna.gr/webServer)

is dedicated to miRNA target prediction/functional analysis, and it is being

widely used from the scientific community, since its initial launch in 2009.

DIANA-microT v5.0, the new version of the microT server, has been significantly

enhanced with an improved target prediction algorithm, DIANA-microT-CDS. It has

been updated to incorporate miRBase version 18 and Ensembl version 69. The in

silico-predicted miRNA-gene interactions in Homo sapiens, Mus musculus,

Drosophila melanogaster and Caenorhabditis elegans exceed 11 million in total.

The web server was completely redesigned, to host a series of sophisticated

workflows, which can be used directly from the on-line web interface, enabling

users without the necessary bioinformatics infrastructure to perform advanced

multi-step functional miRNA analyses. For instance, one available pipeline

performs miRNA target prediction using different thresholds and meta-analysis

statistics, followed by pathway enrichment analysis. DIANA-microT web server v5.0

also supports a complete integration with the Taverna Workflow Management System

(WMS), using the in-house developed DIANA-Taverna Plug-in. This plug-in provides

ready-to-use modules for miRNA target prediction and functional analysis, which

can be used to form advanced high-throughput analysis pipelines.

DOI: 10.1093/nar/gkt393

PMCID: PMC3692048

PMID: 23680784 [Indexed for MEDLINE]

1113. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W329-32. doi:

10.1093/nar/gkt406. Epub 2013 May 15.

CovalentDock Cloud: a web server for automated covalent docking.

Ouyang X(1), Zhou S, Ge Z, Li R, Kwoh CK.

Author information:

(1)State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical

Sciences, Peking University, Beijing, 100191, China.

Covalent binding is an important mechanism for many drugs to gain its function.

We developed a computational algorithm to model this chemical event and extended

it to a web server, the CovalentDock Cloud, to make it accessible directly online

without any local installation and configuration. It provides a simple yet

user-friendly web interface to perform covalent docking experiments and analysis

online. The web server accepts the structures of both the ligand and the receptor

uploaded by the user or retrieved from online databases with valid access id. It

identifies the potential covalent binding patterns, carries out the covalent

docking experiments and provides visualization of the result for user analysis.

This web server is free and open to all users at http://docking.sce.ntu.edu.sg/.

DOI: 10.1093/nar/gkt406

PMCID: PMC3692115

PMID: 23677616 [Indexed for MEDLINE]

1114. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W22-8. doi: 10.1093/nar/gkt389.

Epub 2013 May 15.

aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT

sequence alignment server with enhanced interactivity.

Kuraku S(1), Zmasek CM, Nishimura O, Katoh K.

Author information:

(1)Genome Resource and Analysis Unit, RIKEN Center for Developmental Biology,

Kobe, Hyogo 650-0047, Japan. shigehiro-kuraku@cdb.riken.jp

We report a new web server, aLeaves (http://aleaves.cdb.riken.jp/), for homologue

collection from diverse animal genomes. In molecular comparative studies

involving multiple species, orthology identification is the basis on which most

subsequent biological analyses rely. It can be achieved most accurately by

explicit phylogenetic inference. More and more species are subjected to

large-scale sequencing, but the resultant resources are scattered in independent

project-based, and multi-species, but separate, web sites. This complicates data

access and is becoming a serious barrier to the comprehensiveness of molecular

phylogenetic analysis. aLeaves, launched to overcome this difficulty, collects

sequences similar to an input query sequence from various data sources. The

collected sequences can be passed on to the MAFFT sequence alignment server

(http://mafft.cbrc.jp/alignment/server/), which has been significantly improved

in interactivity. This update enables to switch between (i) sequence selection

using the Archaeopteryx tree viewer, (ii) multiple sequence alignment and (iii)

tree inference. This can be performed as a loop until one reaches a sensible data

set, which minimizes redundancy for better visibility and handling in

phylogenetic inference while covering relevant taxa. The work flow achieved by

the seamless link between aLeaves and MAFFT provides a convenient online platform

to address various questions in zoology and evolutionary biology.

DOI: 10.1093/nar/gkt389

PMCID: PMC3692103

PMID: 23677614 [Indexed for MEDLINE]

1115. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W504-9. doi: 10.1093/nar/gkt398.

Epub 2013 May 15.

WebScipio: Reconstructing alternative splice variants of eukaryotic proteins.

Hatje K(1), Hammesfahr B, Kollmar M.

Author information:

(1)Group Systems Biology of Motor Proteins, Department of NMR-based Structural

Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen 37077,

Germany.

Accurate exon-intron structures are essential prerequisites in genomics,

proteomics and for many protein family and single gene studies. We originally

developed Scipio and the corresponding web service WebScipio for the

reconstruction of gene structures based on protein sequences and available genome

assemblies. WebScipio also allows predicting mutually exclusive spliced exons and

tandemly arrayed gene duplicates. The obtained gene structures are illustrated in

graphical schemes and can be analysed down to the nucleotide level. The set of

eukaryotic genomes available at the WebScipio server is updated on a daily basis.

The current version of the web server provides access to ∼3400 genome assembly

files of >1100 sequenced eukaryotic species. Here, we have also extended the

functionality by adding a module with which expressed sequence tag (EST) and cDNA

data can be mapped to the reconstructed gene structure for the identification of

all types of alternative splice variants. WebScipio has a user-friendly web

interface, and we believe that the improved web server will provide better

service to biologists interested in the gene structure corresponding to their

protein of interest, including all types of alternative splice forms and tandem

gene duplicates. WebScipio is freely available at http://www.webscipio.org.

DOI: 10.1093/nar/gkt398

PMCID: PMC3692071

PMID: 23677611 [Indexed for MEDLINE]

1116. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W363-7. doi: 10.1093/nar/gkt385.

Epub 2013 May 13.

3DEM Loupe: Analysis of macromolecular dynamics using structures from electron

microscopy.

Nogales-Cadenas R(1), Jonic S, Tama F, Arteni AA, Tabas-Madrid D, Vázquez M,

Pascual-Montano A, Sorzano CO.

Author information:

(1)National Center for Biotechnology-CSIC, Madrid 28049, Spain.

Electron microscopy (EM) provides access to structural information of

macromolecular complexes in the 3-20 Å resolution range. Normal mode analysis has

been extensively used with atomic resolution structures and successfully applied

to EM structures. The major application of normal modes is the identification of

possible conformational changes in proteins. The analysis can throw light on the

mechanism following ligand binding, protein-protein interactions, channel opening

and other functional macromolecular movements. In this article, we present a new

web server, 3DEM Loupe, which allows normal mode analysis of any uploaded EM

volume using a user-friendly interface and an intuitive workflow. Results can be

fully explored in 3D through animations and movies generated by the server. The

application is freely available at http://3demloupe.cnb.csic.es.

DOI: 10.1093/nar/gkt385

PMCID: PMC3692114

PMID: 23671335 [Indexed for MEDLINE]

1117. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W601-6. doi: 10.1093/nar/gkt392.

Epub 2013 May 13.

A new reference implementation of the PSICQUIC web service.

del-Toro N(1), Dumousseau M, Orchard S, Jimenez RC, Galeota E, Launay G, Goll J,

Breuer K, Ono K, Salwinski L, Hermjakob H.

Author information:

(1)EMBL-European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK.

ntoro@ebi.ac.uk

The Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC)

specification was created by the Human Proteome Organization Proteomics Standards

Initiative (HUPO-PSI) to enable computational access to molecular-interaction

data resources by means of a standard Web Service and query language. Currently

providing >150 million binary interaction evidences from 28 servers globally, the

PSICQUIC interface allows the concurrent search of multiple molecular-interaction

information resources using a single query. Here, we present an extension of the

PSICQUIC specification (version 1.3), which has been released to be compliant

with the enhanced standards in molecular interactions. The new release also

includes a new reference implementation of the PSICQUIC server available to the

data providers. It offers augmented web service capabilities and improves the

user experience. PSICQUIC has been running for almost 5 years, with a user base

growing from only 4 data providers to 28 (April 2013) allowing access to 151 310

109 binary interactions. The power of this web service is shown in PSICQUIC View

web application, an example of how to simultaneously query, browse and download

results from the different PSICQUIC servers. This application is free and open to

all users with no login requirement

(http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml).

DOI: 10.1093/nar/gkt392

PMCID: PMC3977660

PMID: 23671334 [Indexed for MEDLINE]

1118. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W297-302. doi:

10.1093/nar/gkt380. Epub 2013 May 13.

MoMA-LigPath: a web server to simulate protein-ligand unbinding.

Devaurs D(1), Bouard L, Vaisset M, Zanon C, Al-Bluwi I, Iehl R, Siméon T, Cortés

J.

Author information:

(1)CNRS, LAAS, 7 av du colonel Roche, F-31400 Toulouse, France.

Protein-ligand interactions taking place far away from the active site, during

ligand binding or release, may determine molecular specificity and activity.

However, obtaining information about these interactions with experimental or

computational methods remains difficult. The computational tool presented in this

article, MoMA-LigPath, is based on a mechanistic representation of the molecular

system, considering partial flexibility, and on the application of a

robotics-inspired algorithm to explore the conformational space. Such a purely

geometric approach, together with the efficiency of the exploration algorithm,

enables the simulation of ligand unbinding within short computing time. Ligand

unbinding pathways generated by MoMA-LigPath are a first approximation that can

provide useful information about protein-ligand interactions. When needed, this

approximation can be subsequently refined and analyzed using state-of-the-art

energy models and molecular modeling methods. MoMA-LigPath is available at

http://moma.laas.fr. The web server is free and open to all users, with no login

requirement.

DOI: 10.1093/nar/gkt380

PMCID: PMC3692135

PMID: 23671332 [Indexed for MEDLINE]

1119. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W273-80. doi:

10.1093/nar/gkt384. Epub 2013 May 13.

EvoDesign: De novo protein design based on structural and evolutionary profiles.

Mitra P(1), Shultis D, Zhang Y.

Author information:

(1)Department of Computational Medicine and Bioinformatics, University of

Michigan, Ann Arbor, MI 48109 USA.

Protein design aims to identify new protein sequences of desirable structure and

biological function. Most current de novo protein design methods rely on

physics-based force fields to search for low free-energy states following

Anfinsen's thermodynamic hypothesis. A major obstacle of such approaches is the

inaccuracy of the force field design, which cannot accurately describe the atomic

interactions or distinguish correct folds. We developed a new web server,

EvoDesign, to design optimal protein sequences of given scaffolds along with

multiple sequence and structure-based features to assess the foldability and

goodness of the designs. EvoDesign uses an evolution-profile-based Monte Carlo

search with the profiles constructed from homologous structure families in the

Protein Data Bank. A set of local structure features, including secondary

structure, torsion angle and solvation, are predicted by single-sequence

neural-network training and used to smooth the sequence motif and accommodate the

physicochemical packing. The EvoDesign algorithm has been extensively tested in

large-scale protein design experiments, which demonstrate enhanced foldability

and structural stability of designed sequences compared with the physics-based

designing methods. The EvoDesign server is freely available at

http://zhanglab.ccmb.med.umich.edu/EvoDesign.

DOI: 10.1093/nar/gkt384

PMCID: PMC3692067

PMID: 23671331 [Indexed for MEDLINE]

1120. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W89-97. doi: 10.1093/nar/gkt386.

Epub 2013 May 10.

Graphite Web: Web tool for gene set analysis exploiting pathway topology.

Sales G(1), Calura E, Martini P, Romualdi C.

Author information:

(1)Department of Biology, University of Padova, Via U. Bassi 58/B, 35121 Padova,

Italy.

Graphite web is a novel web tool for pathway analyses and network visualization

for gene expression data of both microarray and RNA-seq experiments. Several

pathway analyses have been proposed either in the univariate or in the global and

multivariate context to tackle the complexity and the interpretation of

expression results. These methods can be further divided into 'topological' and

'non-topological' methods according to their ability to gain power from pathway

topology. Biological pathways are, in fact, not only gene lists but can be

represented through a network where genes and connections are, respectively,

nodes and edges. To this day, the most used approaches are non-topological and

univariate although they miss the relationship among genes. On the contrary,

topological and multivariate approaches are more powerful, but difficult to be

used by researchers without bioinformatic skills. Here we present Graphite web,

the first public web server for pathway analysis on gene expression data that

combines topological and multivariate pathway analyses with an efficient system

of interactive network visualizations for easy results interpretation.

Specifically, Graphite web implements five different gene set analyses on three

model organisms and two pathway databases. Graphite Web is freely available at

http://graphiteweb.bio.unipd.it/.

DOI: 10.1093/nar/gkt386

PMCID: PMC3977659

PMID: 23666626 [Indexed for MEDLINE]

1121. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W238-41. doi:

10.1093/nar/gkt377. Epub 2013 May 9.

STRAW: Species TRee Analysis Web server.

Shaw TI(1), Ruan Z, Glenn TC, Liu L.

Author information:

(1)Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA.

The coalescent methods for species tree reconstruction are increasingly popular

because they can accommodate coalescence and multilocus data sets. Herein, we

present STRAW, a web server that offers workflows for reconstruction of

phylogenies of species using three species tree methods-MP-EST, STAR and NJst.

The input data are a collection of rooted gene trees (for STAR and MP-EST

methods) or unrooted gene trees (for NJst). The output includes the estimated

species tree, modified Robinson-Foulds distances between gene trees and the

estimated species tree and visualization of trees to compare gene trees with the

estimated species tree. The web sever is available at

http://bioinformatics.publichealth.uga.edu/SpeciesTreeAnalysis/.

DOI: 10.1093/nar/gkt377

PMCID: PMC3692081

PMID: 23661681 [Indexed for MEDLINE]

1122. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W427-31. doi:

10.1093/nar/gkt332. Epub 2013 May 8.

CABS-flex: Server for fast simulation of protein structure fluctuations.

Jamroz M(1), Kolinski A, Kmiecik S.

Author information:

(1)Faculty of Chemistry, University of Warsaw, Warsaw 02-093, Poland.

The CABS-flex server (http://biocomp.chem.uw.edu.pl/CABSflex) implements

CABS-model-based protocol for the fast simulations of near-native dynamics of

globular proteins. In this application, the CABS model was shown to be a

computationally efficient alternative to all-atom molecular dynamics--a classical

simulation approach. The simulation method has been validated on a large set of

molecular dynamics simulation data. Using a single input (user-provided file in

PDB format), the CABS-flex server outputs an ensemble of protein models (in

all-atom PDB format) reflecting the flexibility of the input structure, together

with the accompanying analysis (residue mean-square-fluctuation profile and

others). The ensemble of predicted models can be used in structure-based studies

of protein functions and interactions.

DOI: 10.1093/nar/gkt332

PMCID: PMC3692091

PMID: 23658222 [Indexed for MEDLINE]

1123. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W557-61. doi:

10.1093/nar/gkt328. Epub 2013 May 2.

The Taverna workflow suite: designing and executing workflows of Web Services on

the desktop, web or in the cloud.

Wolstencroft K(1), Haines R, Fellows D, Williams A, Withers D, Owen S,

Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, Bhagat J, Belhajjame K, Bacall F,

Hardisty A, Nieva de la Hidalga A, Balcazar Vargas MP, Sufi S, Goble C.

Author information:

(1)School of Computer Science, University of Manchester, Manchester, M13 9PL, UK.

katherine.wolstencroft@manchester.ac.uk

The Taverna workflow tool suite (http://www.taverna.org.uk) is designed to

combine distributed Web Services and/or local tools into complex analysis

pipelines. These pipelines can be executed on local desktop machines or through

larger infrastructure (such as supercomputers, Grids or cloud environments),

using the Taverna Server. In bioinformatics, Taverna workflows are typically used

in the areas of high-throughput omics analyses (for example, proteomics or

transcriptomics), or for evidence gathering methods involving text mining or data

mining. Through Taverna, scientists have access to several thousand different

tools and resources that are freely available from a large range of life science

institutions. Once constructed, the workflows are reusable, executable

bioinformatics protocols that can be shared, reused and repurposed. A repository

of public workflows is available at http://www.myexperiment.org. This article

provides an update to the Taverna tool suite, highlighting new features and

developments in the workbench and the Taverna Server.

DOI: 10.1093/nar/gkt328

PMCID: PMC3692062

PMID: 23640334 [Indexed for MEDLINE]

1124. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W379-83. doi:

10.1093/nar/gkt331. Epub 2013 May 2.

Memoir: template-based structure prediction for membrane proteins.

Ebejer JP(1), Hill JR, Kelm S, Shi J, Deane CM.

Author information:

(1)Department of Statistics, Oxford University, Oxford, OX1 3TG, UK.

Membrane proteins are estimated to be the targets of 50% of drugs that are

currently in development, yet we have few membrane protein crystal structures. As

a result, for a membrane protein of interest, the much-needed structural

information usually comes from a homology model. Current homology modelling

software is optimized for globular proteins, and ignores the constraints that the

membrane is known to place on protein structure. Our Memoir server produces

homology models using alignment and coordinate generation software that has been

designed specifically for transmembrane proteins. Memoir is easy to use, with the

only inputs being a structural template and the sequence that is to be modelled.

We provide a video tutorial and a guide to assessing model quality. Supporting

data aid manual refinement of the models. These data include a set of alternative

conformations for each modelled loop, and a multiple sequence alignment that

incorporates the query and template. Memoir works with both α-helical and

β-barrel types of membrane proteins and is freely available at

http://opig.stats.ox.ac.uk/webapps/memoir.

DOI: 10.1093/nar/gkt331

PMCID: PMC3692111

PMID: 23640332 [Indexed for MEDLINE]

1125. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W133-41. doi:

10.1093/nar/gkt342. Epub 2013 Apr 30.

The Genomic HyperBrowser: an analysis web server for genome-scale data.

Sandve GK(1), Gundersen S, Johansen M, Glad IK, Gunathasan K, Holden L, Holden M,

Liestøl K, Nygård S, Nygaard V, Paulsen J, Rydbeck H, Trengereid K, Clancy T,

Drabløs F, Ferkingstad E, Kalas M, Lien T, Rye MB, Frigessi A, Hovig E.

Author information:

(1)Department of Informatics, University of Oslo, PO Box 1080, Blindern, 0316

Oslo, Norway.

The immense increase in availability of genomic scale datasets, such as those

provided by the ENCODE and Roadmap Epigenomics projects, presents unprecedented

opportunities for individual researchers to pose novel falsifiable biological

questions. With this opportunity, however, researchers are faced with the

challenge of how to best analyze and interpret their genome-scale datasets. A

powerful way of representing genome-scale data is as feature-specific coordinates

relative to reference genome assemblies, i.e. as genomic tracks. The Genomic

HyperBrowser (http://hyperbrowser.uio.no) is an open-ended web server for the

analysis of genomic track data. Through the provision of several highly

customizable components for processing and statistical analysis of genomic

tracks, the HyperBrowser opens for a range of genomic investigations, related to,

e.g., gene regulation, disease association or epigenetic modifications of the

genome.

DOI: 10.1093/nar/gkt342

PMCID: PMC3692097

PMID: 23632163 [Indexed for MEDLINE]

1126. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W98-103. doi:

10.1093/nar/gkt281. Epub 2013 Apr 30.

PlantGSEA: a gene set enrichment analysis toolkit for plant community.

Yi X(1), Du Z, Su Z.

Author information:

(1)State Key Laboratory of Plant Physiology and Biochemistry, College of

Biological Sciences, China Agricultural University, Beijing 100193, China.

Gene Set Enrichment Analysis (GSEA) is a powerful method for interpreting

biological meaning of a list of genes by computing the overlaps with various

previously defined gene sets. As one of the most widely used annotations for

defining gene sets, Gene Ontology (GO) system has been used in many enrichment

analysis tools. EasyGO and agriGO, two GO enrichment analysis toolkits developed

by our laboratory, have gained extensive usage and citations since their releases

because of their effective performance and consistent maintenance. Responding to

the increasing demands of more comprehensive analysis from the users, we

developed a web server as an important component of our bioinformatics analysis

toolkit, named PlantGSEA, which is based on GSEA method and mainly focuses on

plant organisms. In PlantGSEA, 20 290 defined gene sets deriving from different

resources were collected and used for GSEA analysis. The PlantGSEA currently

supports gene locus IDs and Affymatrix microarray probe set IDs from four plant

model species (Arabidopsis thaliana, Oryza sativa, Zea mays and Gossypium

raimondii). The PlantGSEA is an efficient and user-friendly web server, and now

it is publicly accessible at http://structuralbiology.cau.edu.cn/PlantGSEA.

DOI: 10.1093/nar/gkt281

PMCID: PMC3692080

PMID: 23632162 [Indexed for MEDLINE]

1127. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W475-9. doi: 10.1093/nar/gkt291.

Epub 2013 Apr 29.

The RNAsnp web server: predicting SNP effects on local RNA secondary structure.

Sabarinathan R(1), Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J.

Author information:

(1)Center for non-coding RNA in Technology and Health, University of Copenhagen,

Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark.

The function of many non-coding RNA genes and cis-regulatory elements of

messenger RNA largely depends on the structure, which is in turn determined by

their sequence. Single nucleotide polymorphisms (SNPs) and other mutations may

disrupt the RNA structure, interfere with the molecular function and hence cause

a phenotypic effect. RNAsnp is an efficient method to predict the effect of SNPs

on local RNA secondary structure based on the RNA folding algorithms implemented

in the Vienna RNA package. The SNP effects are quantified in terms of empirical

P-values, which, for computational efficiency, are derived from extensive

pre-computed tables of distributions of substitution effects as a function of

gene length and GC content. Here, we present a web service that not only provides

an interface for RNAsnp but also features a graphical output representation. In

addition, the web server is connected to a local mirror of the UCSC genome

browser database that enables the users to select the genomic sequences for

analysis and visualize the results directly in the UCSC genome browser. The

RNAsnp web server is freely available at: http://rth.dk/resources/rnasnp/.

DOI: 10.1093/nar/gkt291

PMCID: PMC3977658

PMID: 23630321 [Indexed for MEDLINE]

1128. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W417-21. doi:

10.1093/nar/gkt287. Epub 2013 Apr 26.

ValiDichro: a website for validating and quality control of protein circular

dichroism spectra.

Woollett B(1), Whitmore L, Janes RW, Wallace BA.

Author information:

(1)Institute of Structural and Molecular Biology, Birkbeck College, University of

London, London WC1E 7HX, UK.

Circular dichroism (CD) spectroscopy is widely used in structural biology as a

technique for examining the structure, folding and conformational changes of

proteins. A new server, ValiDichro, has been developed for checking the quality

and validity of CD spectral data and metadata, both as an aid to data collection

and processing and as a validation procedure for spectra to be included in

publications. ValiDichro currently includes 25 tests for data completeness,

consistency and quality. For each test that is done, not only is a validation

report produced, but the user is also provided with suggestions for correcting or

improving the data. The ValiDichro server is freely available at

http://valispec.cryst.bbk.ac.uk/circularDichroism/ValiDichro/upload.html.

DOI: 10.1093/nar/gkt287

PMCID: PMC3977657

PMID: 23625965 [Indexed for MEDLINE]

1129. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W368-72. doi:

10.1093/nar/gkt294. Epub 2013 Apr 25.

The ModFOLD4 server for the quality assessment of 3D protein models.

McGuffin LJ(1), Buenavista MT, Roche DB.

Author information:

(1)School of Biological Sciences, University of Reading, Reading, RG6 6AS, UK.

l.j.mcguffin@reading.ac.uk

Once you have generated a 3D model of a protein, how do you know whether it bears

any resemblance to the actual structure? To determine the usefulness of 3D models

of proteins, they must be assessed in terms of their quality by methods that

predict their similarity to the native structure. The ModFOLD4 server is the

latest version of our leading independent server for the estimation of both the

global and local (per-residue) quality of 3D protein models. The server produces

both machine readable and graphical output, providing users with intuitive visual

reports on the quality of predicted protein tertiary structures. The ModFOLD4

server is freely available to all at: http://www.reading.ac.uk/bioinf/ModFOLD/.

DOI: 10.1093/nar/gkt294

PMCID: PMC3692122

PMID: 23620298 [Indexed for MEDLINE]

1130. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W3-7. doi: 10.1093/nar/gkt283.

Epub 2013 Apr 24.

DIALIGN at GOBICS--multiple sequence alignment using various sources of external

information.

Al Ait L(1), Yamak Z, Morgenstern B.

Author information:

(1)Department of Bioinformatics, University of Göttingen, Institute of

Microbiology and Genetics, Goldschmidtstr. 1, 37077 Göttingen, Germany.

DIALIGN is an established tool for multiple sequence alignment that is

particularly useful to detect local homologies in sequences with low overall

similarity. In recent years, various versions of the program have been developed,

some of which are fully automated, whereas others are able to accept

user-specified external information. In this article, we review some versions of

the program that are available through 'Göttingen Bioinformatics Compute Server'.

In addition to previously described implementations, we present a new release of

DIALIGN called 'DIALIGN-PFAM', which uses hits to the PFAM database for improved

protein alignment. Our software is available through http://dialign.gobics.de/.

DOI: 10.1093/nar/gkt283

PMCID: PMC3692126

PMID: 23620293 [Indexed for MEDLINE]

1131. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W471-4. doi: 10.1093/nar/gkt290.

Epub 2013 Apr 24.

RNAstructure: Web servers for RNA secondary structure prediction and analysis.

Bellaousov S(1), Reuter JS, Seetin MG, Mathews DH.

Author information:

(1)Department of Biochemistry and Biophysics, University of Rochester Medical

Center, 601 Elmwood Avenue, Box 712, Rochester, NY 14642, USA.

RNAstructure is a software package for RNA secondary structure prediction and

analysis. This contribution describes a new set of web servers to provide its

functionality. The web server offers RNA secondary structure prediction,

including free energy minimization, maximum expected accuracy structure

prediction and pseudoknot prediction. Bimolecular secondary structure prediction

is also provided. Additionally, the server can predict secondary structures

conserved in either two homologs or more than two homologs. Folding free energy

changes can be predicted for a given RNA structure using nearest neighbor rules.

Secondary structures can be compared using circular plots or the scoring methods,

sensitivity and positive predictive value. Additionally, structure drawings can

be rendered as SVG, postscript, jpeg or pdf. The web server is freely available

for public use at: http://rna.urmc.rochester.edu/RNAstructureWeb.

DOI: 10.1093/nar/gkt290

PMCID: PMC3692136

PMID: 23620284 [Indexed for MEDLINE]

1132. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W292-6. doi: 10.1093/nar/gkt300.

Epub 2013 Apr 22.

LISE: a server using ligand-interacting and site-enriched protein triangles for

prediction of ligand-binding sites.

Xie ZR(1), Liu CK, Hsiao FC, Yao A, Hwang MJ.

Author information:

(1)Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan.

LISE is a web server for a novel method for predicting small molecule binding

sites on proteins. It differs from a number of servers currently available for

such predictions in two aspects. First, rather than relying on knowledge of

similar protein structures, identification of surface cavities or estimation of

binding energy, LISE computes a score by counting geometric motifs extracted from

sub-structures of interaction networks connecting protein and ligand atoms. These

network motifs take into account spatial and physicochemical properties of

ligand-interacting protein surface atoms. Second, LISE has now been more

thoroughly tested, as, in addition to the evaluation we previously reported using

two commonly used small benchmark test sets and targets of two community-based

experiments on ligand-binding site predictions, we now report an evaluation using

a large non-redundant data set containing >2000 protein-ligand complexes. This

unprecedented test, the largest ever reported to our knowledge, demonstrates

LISE's overall accuracy and robustness. Furthermore, we have identified some hard

to predict protein classes and provided an estimate of the performance that can

be expected from a state-of-the-art binding site prediction server, such as LISE,

on a proteome scale. The server is freely available at

http://lise.ibms.sinica.edu.tw.

DOI: 10.1093/nar/gkt300

PMCID: PMC3692107

PMID: 23609546 [Indexed for MEDLINE]

1133. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W340-8. doi: 10.1093/nar/gkt292.

Epub 2013 Apr 22.

CNA web server: rigidity theory-based thermal unfolding simulations of proteins

for linking structure, (thermo-)stability, and function.

Krüger DM(1), Rathi PC, Pfleger C, Gohlke H.

Author information:

(1)Computational Pharmaceutical Chemistry Group, Department of Mathematics and

Natural Sciences, Heinrich-Heine-University, 40225 Düsseldorf, Germany.

The Constraint Network Analysis (CNA) web server provides a user-friendly

interface to the CNA approach developed in our laboratory for linking results

from rigidity analyses to biologically relevant characteristics of a biomolecular

structure. The CNA web server provides a refined modeling of thermal unfolding

simulations that considers the temperature dependence of hydrophobic tethers and

computes a set of global and local indices for quantifying biomacromolecular

stability. From the global indices, phase transition points are identified where

the structure switches from a rigid to a floppy state; these phase transition

points can be related to a protein's (thermo-)stability. Structural weak spots

(unfolding nuclei) are automatically identified, too; this knowledge can be

exploited in data-driven protein engineering. The local indices are useful in

linking flexibility and function and to understand the impact of ligand binding

on protein flexibility. The CNA web server robustly handles small-molecule

ligands in general. To overcome issues of sensitivity with respect to the input

structure, the CNA web server allows performing two ensemble-based variants of

thermal unfolding simulations. The web server output is provided as raw data,

plots and/or Jmol representations. The CNA web server, accessible at

http://cpclab.uni-duesseldorf.de/cna or http://www.cnanalysis.de, is free and

open to all users with no login requirement.

DOI: 10.1093/nar/gkt292

PMCID: PMC3692064

PMID: 23609541 [Indexed for MEDLINE]

1134. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W412-6. doi: 10.1093/nar/gkt299.

Epub 2013 Apr 22.

Adepth: New Representation and its implications for atomic depths of

macromolecules.

Chen SW(1), Pellequer JL.

Author information:

(1)13 Avenue de Mayre, F-30200 Bagnols sur Cèze, France.

We applied the signed distance function (SDF) for representing the depths of

atoms in a macromolecule. The calculations of SDF values were performed on grid

points in a rectangular box that accommodates the macromolecule. The depth for an

atom inside the molecule was then obtained as a result of tri-linear

interpolation of SDF values at the nearest grid points surrounding the atom. For

testing the performance of present program Adepth, we have constructed an

artificial molecule whose atomic depths are known as the gold standard for

accuracy assessments. On average, our results showed that Adepth reached an

accuracy of 1.6% at 0.5 Å of grid spacing, whereas the current reference server

DEPTH reached 7.5%. The Adepth program provides both depth and height

representations; it is capable of computing iso-surfaces for atomic depths and

presenting graphical view of macromolecular shape at some distance away from the

surface. Web interface is available at http://biodev.cea.fr/adepth.

DOI: 10.1093/nar/gkt299

PMCID: PMC3692060

PMID: 23609539 [Indexed for MEDLINE]

1135. Proteins. 2013 Jul;81(7):1127-40. doi: 10.1002/prot.24258. Epub 2013 Apr 10.

Simultaneous prediction of protein secondary structure and transmembrane spans.

Leman JK(1), Mueller R, Karakas M, Woetzel N, Meiler J.

Author information:

(1)Department of Chemistry, Vanderbilt University, Nashville, Tennessee; Center

for Structural Biology, Vanderbilt University, Nashville, Tennessee, USA.

Prediction of transmembrane spans and secondary structure from the protein

sequence is generally the first step in the structural characterization of

(membrane) proteins. Preference of a stretch of amino acids in a protein to form

secondary structure and being placed in the membrane are correlated.

Nevertheless, current methods predict either secondary structure or individual

transmembrane states. We introduce a method that simultaneously predicts the

secondary structure and transmembrane spans from the protein sequence. This

approach not only eliminates the necessity to create a consensus prediction from

possibly contradicting outputs of several predictors but bears the potential to

predict conformational switches, i.e., sequence regions that have a high

probability to change for example from a coil conformation in solution to an

α-helical transmembrane state. An artificial neural network was trained on

databases of 177 membrane proteins and 6048 soluble proteins. The output is a 3 ×

3 dimensional probability matrix for each residue in the sequence that combines

three secondary structure types (helix, strand, coil) and three environment types

(membrane core, interface, solution). The prediction accuracies are 70.3% for

nine possible states, 73.2% for three-state secondary structure prediction, and

94.8% for three-state transmembrane span prediction. These accuracies are

comparable to state-of-the-art predictors of secondary structure (e.g., Psipred)

or transmembrane placement (e.g., OCTOPUS). The method is available as web server

and for download at www.meilerlab.org.

Copyright © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24258

PMCID: PMC5064873

PMID: 23349002 [Indexed for MEDLINE]

1136. Telemed J E Health. 2013 Jul;19(7):535-41. doi: 10.1089/tmj.2012.0103.

Implementing DICOM structured reporting in a large-scale telemedicine network.

von Wangenheim A(1), Barcellos CL Jr, Andrade R, de Carlos Back Giuliano I,

Borgatto AF, de Andrade DF.

Author information:

(1)Post-Graduate Program in Computer Science, Federal University of Santa

Catarina, Florianópolis, Santa Catarina, Brazil. awangenh@inf.ufsc.br

INTRODUCTION: Large-scale asynchronous telemedicine networks can offer a unique

opportunity for the acquisition of detailed epidemiological information if the

data are acquired and handled in an appropriate way. In this work, an approach is

presented for the integration of medical reports in the Digital Imaging and

Communications in Medicine (DICOM) Structured Reporting standard in telemedicine

networks using structured vocabularies.

MATERIALS AND METHODS: The use of these structured vocabularies is extended

beyond radiology, and a case study in telecardiology is presented. The approach

was applied in the context of a real-world statewide public telemedicine network;

nowadays on average 470 written electrocardiographic structured reports daily are

being performed. Cardiologists provided more than 220,000 written structured

reports, and these reports are stored into a central database.

RESULTS: This study was performed during a 12-month period, and it was possible

to examine possible associations between a list of co-morbidities and cardiac

risk factors with a diagnosis that indicates the presence of cardiac ischemia,

cardiac injury, or possible necrosis by using DICOM Structured Reporting. Our

application is responsible for coordinating the process of issuance of reports

through various technologies and devices. The system works as a library in an

HTTP server, which accesses information from studies in DICOM format from the

database and from structured vocabularies.

CONCLUSIONS: Results indicate that traceability of morbidity, diagnoses, and

patient clinical information can be achieved, resulting in an efficient data

mining-friendly framework. A multidevice application for Web-based and

smartphone-based platforms showed that it is a viable solution for applying the

DICOM Structured Reporting standard in telemedicine networks.

DOI: 10.1089/tmj.2012.0103

PMCID: PMC3700088

PMID: 23837517 [Indexed for MEDLINE]

1137. PLoS One. 2013 Jun 28;8(6):e67008. doi: 10.1371/journal.pone.0067008. Print 2013.

In silico platform for prediction of N-, O- and C-glycosites in eukaryotic

protein sequences.

Chauhan JS(1), Rao A, Raghava GP.

Author information:

(1)Bioinformatic Centre, Institute of Microbial Technology, Chandigarh, India.

Glycosylation is one of the most abundant and an important post-translational

modification of proteins. Glycosylated proteins (glycoproteins) are involved in

various cellular biological functions like protein folding, cell-cell

interactions, cell recognition and host-pathogen interactions. A large number of

eukaryotic glycoproteins also have therapeutic and potential technology

applications. Therefore, characterization and analysis of glycosites

(glycosylated residues) in these proteins is of great interest to biologists. In

order to cater these needs a number of in silico tools have been developed over

the years, however, a need to get even better prediction tools remains.

Therefore, in this study we have developed a new webserver GlycoEP for more

accurate prediction of N-linked, O-linked and C-linked glycosites in eukaryotic

glycoproteins using two larger datasets, namely, standard and advanced datasets.

In case of standard datasets no two glycosylated proteins are more similar than

40%; advanced datasets are highly non-redundant where no two glycosites' patterns

(as defined in methods) have more than 60% similarity. Further, based on our

results with several algorihtms developed using different machine-learning

techniques, we found Support Vector Machine (SVM) as optimum tool to develop

glycosite prediction models. Accordingly, using our more stringent and

non-redundant advanced datasets, the SVM based models developed in this study

achieved a prediction accuracy of 84.26%, 86.87% and 91.43% with corresponding

MCC of 0.54, 0.20 and 0.78, for N-, O- and C-linked glycosites, respectively. The

best performing models trained on advanced datasets were then implemented as a

user-friendly web server GlycoEP (http://www.imtech.res.in/raghava/glycoep/).

Additionally, this server provides prediction models developed on standard

datasets and allows users to scan sequons in input protein sequences.

DOI: 10.1371/journal.pone.0067008

PMCID: PMC3695939

PMID: 23840574 [Indexed for MEDLINE]

1138. BMC Genomics. 2013 Jun 13;14:397. doi: 10.1186/1471-2164-14-397.

Web-based visual analysis for high-throughput genomics.

Goecks J(1), Eberhard C, Too T; Galaxy Team, Nekrutenko A, Taylor J.

Author information:

(1)Department of Biology, Emory University, 1510 Clifton Road NE, Atlanta, GA

30322, USA.

BACKGROUND: Visualization plays an essential role in genomics research by making

it possible to observe correlations and trends in large datasets as well as

communicate findings to others. Visual analysis, which combines visualization

with analysis tools to enable seamless use of both approaches for scientific

investigation, offers a powerful method for performing complex genomic analyses.

However, there are numerous challenges that arise when creating rich, interactive

Web-based visualizations/visual analysis applications for high-throughput

genomics. These challenges include managing data flow from Web server to Web

browser, integrating analysis tools and visualizations, and sharing

visualizations with colleagues.

RESULTS: We have created a platform simplifies the creation of Web-based

visualization/visual analysis applications for high-throughput genomics. This

platform provides components that make it simple to efficiently query very large

datasets, draw common representations of genomic data, integrate with analysis

tools, and share or publish fully interactive visualizations. Using this

platform, we have created a Circos-style genome-wide viewer, a generic scatter

plot for correlation analysis, an interactive phylogenetic tree, a scalable

genome browser for next-generation sequencing data, and an application for

systematically exploring tool parameter spaces to find good parameter values. All

visualizations are interactive and fully customizable. The platform is integrated

with the Galaxy (http://galaxyproject.org) genomics workbench, making it easy to

integrate new visual applications into Galaxy.

CONCLUSIONS: Visualization and visual analysis play an important role in

high-throughput genomics experiments, and approaches are needed to make it easier

to create applications for these activities. Our framework provides a foundation

for creating Web-based visualizations and integrating them into Galaxy. Finally,

the visualizations we have created using the framework are useful tools for

high-throughput genomics experiments.

DOI: 10.1186/1471-2164-14-397

PMCID: PMC3691752

PMID: 23758618 [Indexed for MEDLINE]

1139. PLoS One. 2013 Jun 13;8(6):e65558. doi: 10.1371/journal.pone.0065558. Print 2013.

Structural modeling and in silico analysis of human superoxide dismutase 2.

de Carvalho MD(1), De Mesquita JF.

Author information:

(1)Bioinformatics and Computational Biology Group, Department of Genetics and

Molecular Biology, Federal University of Rio de Janeiro State, Rio de Janeiro,

Brazil.

Aging in the world population has increased every year. Superoxide dismutase 2

(Mn-SOD or SOD2) protects against oxidative stress, a main factor influencing

cellular longevity. Polymorphisms in SOD2 have been associated with the

development of neurodegenerative diseases, such as Alzheimer's and Parkinson's

disease, as well as psychiatric disorders, such as schizophrenia, depression and

bipolar disorder. In this study, all of the described natural variants (S10I,

A16V, E66V, G76R, I82T and R156W) of SOD2 were subjected to in silico analysis

using eight different algorithms: SNPeffect, PolyPhen-2, PhD-SNP, PMUT, SIFT,

SNAP, SNPs&GO and nsSNPAnalyzer. This analysis revealed disparate results for a

few of the algorithms. The results showed that, from at least one algorithm, each

amino acid substitution appears to harmfully affect the protein. Structural

theoretical models were created for variants through comparative modelling

performed using the MHOLline server (which includes MODELLER and PROCHECK) and ab

initio modelling, using the I-Tasser server. The predicted models were evaluated

using TM-align, and the results show that the models were constructed with high

accuracy. The RMSD values of the modelled mutants indicated likely pathogenicity

for all missense mutations. Structural phylogenetic analysis using ConSurf

revealed that human SOD2 is highly conserved. As a result, a human-curated

database was generated that enables biologists and clinicians to explore SOD2

nsSNPs, including predictions of their effects and visualisation of the alignment

of both the wild-type and mutant structures. The database is freely available at

http://bioinfogroup.com/database and will be regularly updated.

DOI: 10.1371/journal.pone.0065558

PMCID: PMC3681941

PMID: 23785434 [Indexed for MEDLINE]

1140. BMC Res Notes. 2013 Jun 11;6:227. doi: 10.1186/1756-0500-6-227.

Monitoring the antigenic evolution of human influenza A viruses to understand how

and when viruses escape from existing immunity.

Liao YC(1), Lin HH, Lin CH.

Author information:

(1)Division of Biostatistics and Bioinformatics, Institute of Population Health

Sciences, National Health Research Institutes, Zhunan 350, Taiwan.

jade@nhri.org.tw

BACKGROUND: The World Health Organization (WHO) organizes consultations in

February and September of each year, spearheaded by an advisory group of experts

to analyze influenza surveillance data generated by the WHO Global Influenza

Surveillance and Response System (GISRS). The purpose of these consultations is

to recommend the composition on influenza virus vaccines for the northern and

southern hemispheres, respectively. The latest news of influenza viruses is made

available to the public and updated on the WHO website. Although WHO discloses

the manner in which it has made the recommendation, usually by considering

epidemiological and clinical information to analyze the antigenic and genetic

characteristics of seasonal influenza viruses, most individuals do not possess an

understanding of antigenic drift and when it occurs.

FINDINGS: We have constructed a web server, named Fluctrl, and implemented a

pipeline whereby HA sequence data is downloaded from the Influenza Virus Resource

at NCBI along with their isolation information including isolation year and

location, which are parsed and managed in MySQL database. By analyzing the

frequency of each amino acid residue of the HA1 domain expressed by the viruses

on annual basis, users are able to obtain evolutionary dynamics of human

influenza viruses corresponding with epidemics. Users are able to upload and

analyze their HA1 sequences for generating evolutionary dynamics. In addition, a

distribution of amino acid residues at a particular site is represented

geographically to trace the location where antigenic variants are seeded.

CONCLUSIONS: Fluctrl is constructed for monitoring the antigenic evolution of

human influenza A viruses. This tool is intended to inform the general public how

and when influenza viruses evade the human body's immunity. Furthermore,

leveraging the geographic information, the original locations of emerging

influenza viruses can be traced. Fluctrl is freely accessible at

http://sb.nhri.org.tw/fluctrl.

DOI: 10.1186/1756-0500-6-227

PMCID: PMC3689074

PMID: 23758844 [Indexed for MEDLINE]

1141. Acta Biotheor. 2013 Jun;61(2):259-68. doi: 10.1007/s10441-013-9181-9. Epub 2013

Mar 10.

Using over-represented tetrapeptides to predict protein submitochondria

locations.

Lin H(1), Chen W, Yuan LF, Li ZQ, Ding H.

Author information:

(1)Key Laboratory for NeuroInformation of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu, China. hlin@uestc.edu.cn

The mitochondrion is a key organelle of eukaryotic cell that provides the energy

for cellular activities. Correctly identifying submitochondria locations of

proteins can provide plentiful information for understanding their functions.

However, using web-experimental methods to recognize submitochondria locations of

proteins are time-consuming and costly. Thus, it is highly desired to develop a

bioinformatics method to predict the submitochondria locations of mitochondrion

proteins. In this work, a novel method based on support vector machine was

developed to predict the submitochondria locations of mitochondrion proteins by

using over-represented tetrapeptides selected by using binomial distribution. A

reliable and rigorous benchmark dataset including 495 mitochondrion proteins with

sequence identity ≤25% was constructed for testing and evaluating the proposed

model. Jackknife cross-validated results showed that the 91.1% of the 495

mitochondrion proteins can be correctly predicted. Subsequently, our model was

estimated by three existing benchmark datasets. The overall accuracies are 94.0,

94.7 and 93.4%, respectively, suggesting that the proposed model is potentially

useful in the realm of mitochondrion proteome research. Based on this model, we

built a predictor called TetraMito which is freely available at

http://lin.uestc.edu.cn/server/TetraMito.

DOI: 10.1007/s10441-013-9181-9

PMID: 23475502 [Indexed for MEDLINE]

1142. Bioinformatics. 2013 Jun 1;29(11):1375-81. doi: 10.1093/bioinformatics/btt168.

Epub 2013 Apr 24.

A novel web server predicts amino acid residue protection against

hydrogen-deuterium exchange.

Lobanov MY(1), Suvorina MY, Dovidchenko NV, Sokolovskiy IV, Surin AK, Galzitskaya

OV.

Author information:

(1)Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow

Region, Russia.

MOTIVATION: To clarify the relationship between structural elements and

polypeptide chain mobility, a set of statistical analyses of structures is

necessary. Because at present proteins with determined spatial structures are

much less numerous than those with amino acid sequence known, it is important to

be able to predict the extent of proton protection from hydrogen-deuterium (HD)

exchange basing solely on the protein primary structure.

RESULTS: Here we present a novel web server aimed to predict the degree of amino

acid residue protection against HD exchange solely from the primary structure of

the protein chain under study. On the basis of the amino acid sequence, the

presented server offers the following three possibilities (predictors) for user's

choice. First, prediction of the number of contacts occurring in this protein,

which is shown to be helpful in estimating the number of protons protected

against HD exchange (sensitivity 0.71). Second, probability of H-bonding in this

protein, which is useful for finding the number of unprotected protons

(specificity 0.71). The last is the use of an artificial predictor. Also, we

report on mass spectrometry analysis of HD exchange that has been first applied

to free amino acids. Its results showed a good agreement with theoretical data

(number of protons) for 10 globular proteins (correlation coefficient 0.73). We

pioneered in compiling two datasets of experimental HD exchange data for 35

proteins.

AVAILABILITY: The H-Protection server is available for users at

http://bioinfo.protres.ru/ogp/

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt168

PMID: 23620358 [Indexed for MEDLINE]

1143. Bioinformatics. 2013 Jun 1;29(11):1467-8. doi: 10.1093/bioinformatics/btt159.

Epub 2013 Apr 10.

Biographer: web-based editing and rendering of SBGN compliant biochemical

networks.

Krause F(1), Schulz M, Ripkens B, Flöttmann M, Krantz M, Klipp E, Handorf T.

Author information:

(1)Theoretical Biophysics, Humboldt-Universität zu Berlin, Berlin, Germany.

MOTIVATION: The rapid accumulation of knowledge in the field of Systems Biology

during the past years requires advanced, but simple-to-use, methods for the

visualization of information in a structured and easily comprehensible manner.

RESULTS: We have developed biographer, a web-based renderer and editor for

reaction networks, which can be integrated as a library into tools dealing with

network-related information. Our software enables visualizations based on the

emerging standard Systems Biology Graphical Notation. It is able to import

networks encoded in various formats such as SBML, SBGN-ML and jSBGN, a custom

lightweight exchange format. The core package is implemented in HTML5, CSS and

JavaScript and can be used within any kind of web-based project. It features

interactive graph-editing tools and automatic graph layout algorithms. In

addition, we provide a standalone graph editor and a web server, which contains

enhanced features like web services for the import and export of models and

visualizations in different formats.

AVAILABILITY: The biographer tool can be used at and downloaded from the web page

http://biographer.biologie.hu-berlin.de/. The different software packages,

including a server-independent version as well as a web server for Windows and

Linux based systems, are available at http://code.google.com/p/biographer/ under

the open-source license LGPL

DOI: 10.1093/bioinformatics/btt159

PMCID: PMC3661053

PMID: 23574737 [Indexed for MEDLINE]

1144. DNA Res. 2013 Jun;20(3):255-62. doi: 10.1093/dnares/dst007. Epub 2013 Mar 29.

Development and characterization of cDNA resources for the common marmoset: one

of the experimental primate models.

Tatsumoto S(1), Adati N, Tohtoki Y, Sakaki Y, Boroviak T, Habu S, Okano H,

Suemizu H, Sasaki E, Satake M.

Author information:

(1)RIKEN Genomic Sciences Center, Suehiro-cho 1-7-22, Tsurumi-ku, Yokohama

230-0045, Japan.

The common marmoset is a new world monkey, which has become a valuable

experimental animal for biomedical research. This study developed cDNA libraries

for the common marmoset from five different tissues. A total of 290 426

high-quality EST sequences were obtained, where 251 587 sequences (86.5%) had

homology (1E(-100)) with the Refseqs of six different primate species, including

human and marmoset. In parallel, 270 673 sequences (93.2%) were aligned to the

human genome. When 247 090 sequences were assembled into 17 232 contigs, most of

the sequences (218 857 or 15 089 contigs) were located in exonic regions,

indicating that these genes are expressed in human and marmoset. The other 5578

sequences (or 808 contigs) mapping to the human genome were not located in exonic

regions, suggesting that they are not expressed in human. Furthermore, a

different set of 118 potential coding sequences were not similar to any Refseqs

in any species, and, thus, may represent unknown genes. The cDNA libraries

developed in this study are available through RIKEN Bio Resource Center. A Web

server for the marmoset cDNAs is available at

http://marmoset.nig.ac.jp/index.html, where each marmoset EST sequence has been

annotated by reference to the human genome. These new libraries will be a useful

genetic resource to facilitate research in the common marmoset.

DOI: 10.1093/dnares/dst007

PMCID: PMC3686431

PMID: 23543116 [Indexed for MEDLINE]

1145. Fungal Genet Biol. 2013 Jun;55:77-84. doi: 10.1016/j.fgb.2012.09.012. Epub 2012

Oct 31.

Two-component signal transduction in Agaricus bisporus: a comparative genomic

analysis with other basidiomycetes through the web-based tool BASID2CS.

Lavín JL(1), García-Yoldi A, Ramírez L, Pisabarro AG, Oguiza JA.

Author information:

(1)Genetics and Microbiology Research Group, Departamento de Producción Agraria,

Universidad Pública de Navarra, 31006 Pamplona, Spain.

Two-component systems (TCSs) are signal transduction mechanisms present in many

eukaryotes, including fungi that play essential roles in the regulation of

several cellular functions and responses. In this study, we carry out a genomic

analysis of the TCS proteins in two varieties of the white button mushroom

Agaricus bisporus. The genomes of both A. bisporus varieties contain eight genes

coding for TCS proteins, which include four hybrid Histidine Kinases (HKs), a

single histidine-containing phosphotransfer (HPt) protein and three Response

Regulators (RRs). Comparison of the TCS proteins among A. bisporus and the

sequenced basidiomycetes showed a conserved core complement of five TCS proteins

including the Tco1/Nik1 hybrid HK, HPt protein and Ssk1, Skn7 and Rim15-like RRs.

In addition, Dual-HKs, unusual hybrid HKs with 2 HK and 2 RR domains, are absent

in A. bisporus and are limited to various species of basidiomycetes. Differential

expression analysis showed no significant up- or down-regulation of the Agaricus

TCS genes in the conditions/tissue analyzed with the exception of the Skn7-like

RR gene (Agabi\_varbisH97\_2|198669) that is significantly up-regulated on compost

compared to cultured mycelia. Furthermore, the pipeline web server BASID2CS

(http://bioinformatics.unavarra.es:1000/B2CS/BASID2CS.htm) has been specifically

designed for the identification, classification and functional annotation of

putative TCS proteins from any predicted proteome of basidiomycetes using a

combination of several bioinformatic approaches.

Copyright © 2012 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.fgb.2012.09.012

PMID: 23123423 [Indexed for MEDLINE]

1146. J Bacteriol. 2013 Jun;195(11):2463-73. doi: 10.1128/JB.00140-13. Epub 2013 Mar

15.

Genomic reconstruction of the transcriptional regulatory network in Bacillus

subtilis.

Leyn SA(1), Kazanov MD, Sernova NV, Ermakova EO, Novichkov PS, Rodionov DA.

Author information:

(1)Sanford-Burnham Medical Research Institute, La Jolla, California, USA.

The adaptation of microorganisms to their environment is controlled by complex

transcriptional regulatory networks (TRNs), which are still only partially

understood even for model species. Genome scale annotation of regulatory features

of genes and TRN reconstruction are challenging tasks of microbial genomics. We

used the knowledge-driven comparative-genomics approach implemented in the

RegPredict Web server to infer TRN in the model Gram-positive bacterium Bacillus

subtilis and 10 related Bacillales species. For transcription factor (TF)

regulons, we combined the available information from the DBTBS database and the

literature with bioinformatics tools, allowing inference of TF binding sites

(TFBSs), comparative analysis of the genomic context of predicted TFBSs,

functional assignment of target genes, and effector prediction. For RNA regulons,

we used known RNA regulatory motifs collected in the Rfam database to scan

genomes and analyze the genomic context of new RNA sites. The inferred TRN in B.

subtilis comprises regulons for 129 TFs and 24 regulatory RNA families. First, we

analyzed 66 TF regulons with previously known TFBSs in B. subtilis and projected

them to other Bacillales genomes, resulting in refinement of TFBS motifs and

identification of novel regulon members. Second, we inferred motifs and described

regulons for 28 experimentally studied TFs with previously unknown TFBSs. Third,

we discovered novel motifs and reconstructed regulons for 36 previously

uncharacterized TFs. The inferred collection of regulons is available in the

RegPrecise database (http://regprecise.lbl.gov/) and can be used in genetic

experiments, metabolic modeling, and evolutionary analysis.

DOI: 10.1128/JB.00140-13

PMCID: PMC3676070

PMID: 23504016 [Indexed for MEDLINE]

1147. Nucleic Acids Res. 2013 Jun;41(11):e113. doi: 10.1093/nar/gkt234. Epub 2013 Apr

10.

Metabolic tinker: an online tool for guiding the design of synthetic metabolic

pathways.

McClymont K(1), Soyer OS.

Author information:

(1)Computer Science, College of Engineering, Mathematics, and Physical Sciences,

University of Exeter, Exeter EX4 4QF, UK.

One of the primary aims of synthetic biology is to (re)design metabolic pathways

towards the production of desired chemicals. The fast pace of developments in

molecular biology increasingly makes it possible to experimentally redesign

existing pathways and implement de novo ones in microbes or using in vitro

platforms. For such experimental studies, the bottleneck is shifting from

implementation of pathways towards their initial design. Here, we present an

online tool called 'Metabolic Tinker', which aims to guide the design of

synthetic metabolic pathways between any two desired compounds. Given two

user-defined 'target' and 'source' compounds, Metabolic Tinker searches for

thermodynamically feasible paths in the entire known metabolic universe using a

tailored heuristic search strategy. Compared with similar graph-based search

tools, Metabolic Tinker returns a larger number of possible paths owing to its

broad search base and fast heuristic, and provides for the first time

thermodynamic feasibility information for the discovered paths. Metabolic Tinker

is available as a web service at http://osslab.ex.ac.uk/tinker.aspx. The same

website also provides the source code for Metabolic Tinker, allowing it to be

developed further or run on personal machines for specific applications.

DOI: 10.1093/nar/gkt234

PMCID: PMC3675468

PMID: 23580552 [Indexed for MEDLINE]

1148. PLoS One. 2013 May 22;8(5):e63906. doi: 10.1371/journal.pone.0063906. Print 2013.

Serverification of molecular modeling applications: the Rosetta Online Server

that Includes Everyone (ROSIE).

Lyskov S(1), Chou FC, Conchúir SÓ, Der BS, Drew K, Kuroda D, Xu J, Weitzner BD,

Renfrew PD, Sripakdeevong P, Borgo B, Havranek JJ, Kuhlman B, Kortemme T, Bonneau

R, Gray JJ, Das R.

Author information:

(1)Department of Chemical and Biomolecular Engineering, The Johns Hopkins

University, Baltimore, Maryland, USA.

The Rosetta molecular modeling software package provides experimentally tested

and rapidly evolving tools for the 3D structure prediction and high-resolution

design of proteins, nucleic acids, and a growing number of non-natural polymers.

Despite its free availability to academic users and improving documentation, use

of Rosetta has largely remained confined to developers and their immediate

collaborators due to the code's difficulty of use, the requirement for large

computational resources, and the unavailability of servers for most of the

Rosetta applications. Here, we present a unified web framework for Rosetta

applications called ROSIE (Rosetta Online Server that Includes Everyone). ROSIE

provides (a) a common user interface for Rosetta protocols, (b) a stable

application programming interface for developers to add additional protocols, (c)

a flexible back-end to allow leveraging of computer cluster resources shared by

RosettaCommons member institutions, and (d) centralized administration by the

RosettaCommons to ensure continuous maintenance. This paper describes the ROSIE

server infrastructure, a step-by-step 'serverification' protocol for use by

Rosetta developers, and the deployment of the first nine ROSIE applications by

six separate developer teams: Docking, RNA de novo, ERRASER, Antibody, Sequence

Tolerance, Supercharge, Beta peptide design, NCBB design, and VIP redesign. As

illustrated by the number and diversity of these applications, ROSIE offers a

general and speedy paradigm for serverification of Rosetta applications that

incurs negligible cost to developers and lowers barriers to Rosetta use for the

broader biological community. ROSIE is available at

http://rosie.rosettacommons.org.

DOI: 10.1371/journal.pone.0063906

PMCID: PMC3661552

PMID: 23717507 [Indexed for MEDLINE]

1149. Anal Biochem. 2013 May 15;436(2):168-77. doi: 10.1016/j.ab.2013.01.019. Epub 2013

Feb 6.

iAMP-2L: a two-level multi-label classifier for identifying antimicrobial

peptides and their functional types.

Xiao X(1), Wang P, Lin WZ, Jia JH, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China. xxiao@gordonlifescience.org

Antimicrobial peptides (AMPs), also called host defense peptides, are an

evolutionarily conserved component of the innate immune response and are found

among all classes of life. According to their special functions, AMPs are

generally classified into ten categories: Antibacterial Peptides,

Anticancer/tumor Peptides, Antifungal Peptides, Anti-HIV Peptides, Antiviral

Peptides, Antiparasital Peptides, Anti-protist Peptides, AMPs with Chemotactic

Activity, Insecticidal Peptides, and Spermicidal Peptides. Given a query peptide,

how can we identify whether it is an AMP or non-AMP? If it is, can we identify

which functional type or types it belong to? Particularly, how can we deal with

the multi-type problem since an AMP may belong to two or more functional types?

To address these problems, which are obviously very important to both basic

research and drug development, a multi-label classifier was developed based on

the pseudo amino acid composition (PseAAC) and fuzzy K-nearest neighbor (FKNN)

algorithm, where the components of PseAAC were featured by incorporating five

physicochemical properties. The novel classifier is called iAMP-2L, where "2L"

means that it is a 2-level predictor. The 1st-level is to answer the 1st question

above, while the 2nd-level is to answer the 2nd and 3rd questions that are beyond

the reach of any existing methods in this area. For the conveniences of users, a

user-friendly web-server for iAMP-2L was established at

http://www.jci-bioinfo.cn/iAMP-2L.

Copyright © 2013 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2013.01.019

PMID: 23395824 [Indexed for MEDLINE]

1150. BMC Bioinformatics. 2013 May 13;14:158. doi: 10.1186/1471-2105-14-158.

Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient.

Stoltzfus A(1), Lapp H, Matasci N, Deus H, Sidlauskas B, Zmasek CM, Vaidya G,

Pontelli E, Cranston K, Vos R, Webb CO, Harmon LJ, Pirrung M, O'Meara B, Pennell

MW, Mirarab S, Rosenberg MS, Balhoff JP, Bik HM, Heath TA, Midford PE, Brown JW,

McTavish EJ, Sukumaran J, Westneat M, Alfaro ME, Steele A, Jordan G.

Author information:

(1)Institute for Bioscience and Biotechnology Research, Biosystems and

Biomaterials Division, National Institute of Standards and Technology,

Gaithersburg, MD 20899, USA. arlin@umd.edu

BACKGROUND: Scientists rarely reuse expert knowledge of phylogeny, in spite of

years of effort to assemble a great "Tree of Life" (ToL). A notable exception

involves the use of Phylomatic, which provides tools to generate custom

phylogenies from a large, pre-computed, expert phylogeny of plant taxa. This

suggests great potential for a more generalized system that, starting with a

query consisting of a list of any known species, would rectify non-standard

names, identify expert phylogenies containing the implicated taxa, prune away

unneeded parts, and supply branch lengths and annotations, resulting in a custom

phylogeny suited to the user's needs. Such a system could become a sustainable

community resource if implemented as a distributed system of loosely coupled

parts that interact through clearly defined interfaces.

RESULTS: With the aim of building such a "phylotastic" system, the NESCent

Hackathons, Interoperability, Phylogenies (HIP) working group recruited 2 dozen

scientist-programmers to a weeklong programming hackathon in June 2012. During

the hackathon (and a three-month follow-up period), 5 teams produced designs,

implementations, documentation, presentations, and tests including: (1) a

generalized scheme for integrating components; (2) proof-of-concept pruners and

controllers; (3) a meta-API for taxonomic name resolution services; (4) a system

for storing, finding, and retrieving phylogenies using semantic web technologies

for data exchange, storage, and querying; (5) an innovative new service,

DateLife.org, which synthesizes pre-computed, time-calibrated phylogenies to

assign ages to nodes; and (6) demonstration projects. These outcomes are

accessible via a public code repository (GitHub.com), a website

(http://www.phylotastic.org), and a server image.

CONCLUSIONS: Approximately 9 person-months of effort (centered on a software

development hackathon) resulted in the design and implementation of

proof-of-concept software for 4 core phylotastic components, 3 controllers, and 3

end-user demonstration tools. While these products have substantial limitations,

they suggest considerable potential for a distributed system that makes

phylogenetic knowledge readily accessible in computable form. Widespread use of

phylotastic systems will create an electronic marketplace for sharing

phylogenetic knowledge that will spur innovation in other areas of the ToL

enterprise, such as annotation of sources and methods and third-party methods of

quality assessment.

DOI: 10.1186/1471-2105-14-158

PMCID: PMC3669619

PMID: 23668630 [Indexed for MEDLINE]

1151. PLoS One. 2013 May 7;8(5):e62224. doi: 10.1371/journal.pone.0062224. Print 2013.

The PARIGA server for real time filtering and analysis of reciprocal BLAST

results.

Orsini M(1), Carcangiu S, Cuccuru G, Uva P, Tramontano A.

Author information:

(1)CRS4 Bioinformatics Laboratory, Science and Technology Park Polaris, Pula,

Italy. orsini@crs4.it

BLAST-based similarity searches are commonly used in several applications

involving both nucleotide and protein sequences. These applications span from

simple tasks such as mapping sequences over a database to more complex procedures

as clustering or annotation processes. When the amount of analysed data

increases, manual inspection of BLAST results become a tedious procedure. Tools

for parsing or filtering BLAST results for different purposes are then required.

We describe here PARIGA (http://resources.bioinformatica.crs4.it/pariga/), a

server that enables users to perform all-against-all BLAST searches on two sets

of sequences selected by the user. Moreover, since it stores the two BLAST output

in a python-serialized-objects database, results can be filtered according to

several parameters in real-time fashion, without re-running the process and

avoiding additional programming efforts. Results can be interrogated by the user

using logical operations, for example to retrieve cases where two queries match

same targets, or when sequences from the two datasets are reciprocal best hits,

or when a query matches a target in multiple regions. The Pariga web server is

designed to be a helpful tool for managing the results of sequence similarity

searches. The design and implementation of the server renders all operations very

fast and easy to use.

DOI: 10.1371/journal.pone.0062224

PMCID: PMC3646873

PMID: 23667459 [Indexed for MEDLINE]

1152. PLoS One. 2013 May 7;8(5):e62216. doi: 10.1371/journal.pone.0062216. Print 2013.

Improved method for linear B-cell epitope prediction using antigen's primary

sequence.

Singh H(1), Ansari HR, Raghava GP.

Author information:

(1)Bioinformatics Center, Institute of Microbial Technology, Chandigarh, India.

One of the major challenges in designing a peptide-based vaccine is the

identification of antigenic regions in an antigen that can stimulate B-cell's

response, also called B-cell epitopes. In the past, several methods have been

developed for the prediction of conformational and linear (or continuous) B-cell

epitopes. However, the existing methods for predicting linear B-cell epitopes are

far from perfection. In this study, an attempt has been made to develop an

improved method for predicting linear B-cell epitopes. We have retrieved

experimentally validated B-cell epitopes as well as non B-cell epitopes from

Immune Epitope Database and derived two types of datasets called Lbtope\_Variable

and Lbtope\_Fixed length datasets. The Lbtope\_Variable dataset contains 14876

B-cell epitope and 23321 non-epitopes of variable length where as Lbtope\_Fixed

length dataset contains 12063 B-cell epitopes and 20589 non-epitopes of fixed

length. We also evaluated the performance of models on above datasets after

removing highly identical peptides from the datasets. In addition, we have

derived third dataset Lbtope\_Confirm having 1042 epitopes and 1795 non-epitopes

where each epitope or non-epitope has been experimentally validated in at least

two studies. A number of models have been developed to discriminate epitopes and

non-epitopes using different machine-learning techniques like Support Vector

Machine, and K-Nearest Neighbor. We achieved accuracy from ∼54% to 86% using

diverse s features like binary profile, dipeptide composition, AAP (amino acid

pair) profile. In this study, for the first time experimentally validated non

B-cell epitopes have been used for developing method for predicting linear B-cell

epitopes. In previous studies, random peptides have been used as non B-cell

epitopes. In order to provide service to scientific community, a web server

LBtope has been developed for predicting and designing B-cell epitopes

(http://crdd.osdd.net/raghava/lbtope/).

DOI: 10.1371/journal.pone.0062216

PMCID: PMC3646881

PMID: 23667458 [Indexed for MEDLINE]

1153. Amino Acids. 2013 May;44(5):1365-79. doi: 10.1007/s00726-013-1472-6. Epub 2013

Feb 28.

Learning protein multi-view features in complex space.

Yu DJ(1), Hu J, Wu XW, Shen HB, Chen J, Tang ZM, Yang J, Yang JY.

Author information:

(1)School of Computer Science and Engineering, Nanjing University of Science and

Technology, Nanjing 210094, China.

Protein attribute prediction from primary sequences is an important task and how

to extract discriminative features is one of the most crucial aspects. Because

single-view feature cannot reflect all the information of a protein, fusing

multi-view features is considered as a promising route to improve prediction

accuracy. In this paper, we propose a novel framework for protein multi-view

feature fusion: first, features from different views are parallely combined to

form complex feature vectors; Then, we extend the classic principal component

analysis to the generalized principle component analysis for further feature

extraction from the parallely combined complex features, which lie in a complex

space. Finally, the extracted features are used for prediction. Experimental

results on different benchmark datasets and machine learning algorithms

demonstrate that parallel strategy outperforms the traditional serial approach

and is particularly helpful for extracting the core information buried among

multi-view feature sets. A web server for protein structural class prediction

based on the proposed method (COMSPA) is freely available for academic use at:

http://www.csbio.sjtu.edu.cn/bioinf/COMSPA/ .

DOI: 10.1007/s00726-013-1472-6

PMID: 23456487 [Indexed for MEDLINE]

1154. Bioinformatics. 2013 May 1;29(9):1218-9. doi: 10.1093/bioinformatics/btt102. Epub

2013 Mar 7.

FTFlex: accounting for binding site flexibility to improve fragment-based

identification of druggable hot spots.

Grove LE(1), Hall DR, Beglov D, Vajda S, Kozakov D.

Author information:

(1)Department of Sciences, Wentworth Institute of Technology, Boston, MA 01746,

USA.

Computational solvent mapping finds binding hot spots, determines their

druggability and provides information for drug design. While mapping of a

ligand-bound structure yields more accurate results, usually the apo structure

serves as the starting point in design. The FTFlex algorithm, implemented as a

server, can modify an apo structure to yield mapping results that are similar to

those of the respective bound structure. Thus, FTFlex is an extension of our

FTMap server, which only considers rigid structures. FTFlex identifies flexible

residues within the binding site and determines alternative conformations using a

rotamer library. In cases where the mapping results of the apo structure were in

poor agreement with those of the bound structure, FTFlex was able to yield a

modified apo structure, which lead to improved FTMap results. In cases where the

mapping results of the apo and bound structures were in good agreement, no new

structure was predicted.AVAILABILITY: FTFlex is freely available as a web-based

server at http://ftflex.bu.edu/.

DOI: 10.1093/bioinformatics/btt102

PMCID: PMC3634182

PMID: 23476022 [Indexed for MEDLINE]

1155. J Comput Aided Mol Des. 2013 May;27(5):419-26. doi: 10.1007/s10822-013-9654-6.

Epub 2013 May 29.

IS-Dom: a dataset of independent structural domains automatically delineated from

protein structures.

Ebina T(1), Umezawa Y, Kuroda Y.

Author information:

(1)Department of Biotechnology and Life Science, Tokyo University of Agriculture

and Technology, 12-24-16 Nakamachi, Koganei-shi, Tokyo 184-8588, Japan.

teppei-ebina@brain.riken.jp

Protein domains that can fold in isolation are significant targets in diverse

area of proteomics research as they are often readily analyzed by high-throughput

methods. Here, we report IS-Dom, a dataset of Independent Structural Domains

(ISDs) that are most likely to fold in isolation. IS-Dom was constructed by

filtering domains from SCOP, CATH, and DomainParser using quantitative structural

measures, which were calculated by estimating inter-domain hydrophobic clusters

and hydrogen bonds from the full length protein's atomic coordinates. The ISD

detection protocol is fully automated, and all of the computed interactions are

stored in the server which enables rapid update of IS-Dom. We also prepared a

standard IS-Dom using parameters optimized by maximizing the Youden's index. The

standard IS-Dom, contained 54,860 ISDs, of which 25.5 % had high sequence

identity and termini overlap with a Protein Data Bank (PDB) cataloged sequence

and are thus experimentally shown to fold in isolation [coined autonomously

folded domain (AFDs)]. Furthermore, our ISD detection protocol missed less than

10 % of the AFDs, which corroborated our protocol's ability to define structural

domains that are able to fold independently. IS-Dom is available through the web

server ( http://domserv.lab.tuat.ac.jp/IS-Dom.html ), and users can either,

download the standard IS-Dom dataset, construct their own IS-Dom by interactively

varying the parameters, or assess the structural independence of newly defined

putative domains.

DOI: 10.1007/s10822-013-9654-6

PMID: 23715893 [Indexed for MEDLINE]

1156. Mol Biol Evol. 2013 May;30(5):1196-205. doi: 10.1093/molbev/mst030. Epub 2013 Feb

18.

FUBAR: a fast, unconstrained bayesian approximation for inferring selection.

Murrell B(1), Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL,

Scheffler K.

Author information:

(1)Department of Mathematical Sciences, Stellenbosch University, Stellenbosch,

South Africa.

Model-based analyses of natural selection often categorize sites into a

relatively small number of site classes. Forcing each site to belong to one of

these classes places unrealistic constraints on the distribution of selection

parameters, which can result in misleading inference due to model

misspecification. We present an approximate hierarchical Bayesian method using a

Markov chain Monte Carlo (MCMC) routine that ensures robustness against model

misspecification by averaging over a large number of predefined site classes.

This leaves the distribution of selection parameters essentially unconstrained,

and also allows sites experiencing positive and purifying selection to be

identified orders of magnitude faster than by existing methods. We demonstrate

that popular random effects likelihood methods can produce misleading results

when sites assigned to the same site class experience different levels of

positive or purifying selection--an unavoidable scenario when using a small

number of site classes. Our Fast Unconstrained Bayesian AppRoximation (FUBAR) is

unaffected by this problem, while achieving higher power than existing

unconstrained (fixed effects likelihood) methods. The speed advantage of FUBAR

allows us to analyze larger data sets than other methods: We illustrate this on a

large influenza hemagglutinin data set (3,142 sequences). FUBAR is available as a

batch file within the latest HyPhy distribution (http://www.hyphy.org), as well

as on the Datamonkey web server (http://www.datamonkey.org/).

DOI: 10.1093/molbev/mst030

PMCID: PMC3670733

PMID: 23420840 [Indexed for MEDLINE]

1157. Mol Biol Evol. 2013 May;30(5):1032-7. doi: 10.1093/molbev/mst021. Epub 2013 Feb

6.

AGP: a multimethods web server for alignment-free genome phylogeny.

Cheng J(1), Cao F, Liu Z.

Author information:

(1)Nanjing Forestry University, Nanjing, China.

Phylogenetic analysis based on alignment method meets huge challenges when

dealing with whole-genome sequences, for example, recombination, shuffling, and

rearrangement of sequences. Thus, various alignment-free methods for phylogeny

construction have been proposed. However, most of these methods have not been

implemented as tools or web servers. Researchers cannot use these methods easily

with their data sets. To facilitate the usage of various alignment-free methods,

we implemented most of the popular alignment-free methods and constructed a

user-friendly web server for alignment-free genome phylogeny (AGP). AGP

integrated the phylogenetic tree construction, visualization, and comparison

functions together. Both AGP and all source code of the methods are available at

http://www.herbbol.org:8000/agp (last accessed February 26, 2013). AGP will

facilitate research in the field of whole-genome phylogeny and comparison.

DOI: 10.1093/molbev/mst021

PMID: 23389766 [Indexed for MEDLINE]

1158. Proteomics. 2013 May;13(9):1444-56. doi: 10.1002/pmic.201200175.

ESPRESSO: a system for estimating protein expression and solubility in protein

expression systems.

Hirose S(1), Noguchi T.

Author information:

(1)Computational Biology Research Center (CBRC), National Institute of Advanced

Industrial Science and Technology (AIST), Tokyo, Japan. hirose-shuichi@aist.go.jp

Recombinant protein technology is essential for conducting protein science and

using proteins as materials in pharmaceutical or industrial applications.

Although obtaining soluble proteins is still a major experimental obstacle,

knowledge about protein expression/solubility under standard conditions may

increase the efficiency and reduce the cost of proteomics studies. In this study,

we present a computational approach to estimate the probability of protein

expression and solubility for two different protein expression systems: in vivo

Escherichia coli and wheat germ cell-free, from only the sequence information. It

implements two kinds of methods: a sequence/predicted structural property-based

method that uses both the sequence and predicted structural features, and a

sequence pattern-based method that utilizes the occurrence frequencies of

sequence patterns. In the benchmark test, the proposed methods obtained F-scores

of around 70%, and outperformed publicly available servers. Applying the proposed

methods to genomic data revealed that proteins associated with translation or

transcription have a strong tendency to be expressed as soluble proteins by the

in vivo E. coli expression system. The sequence pattern-based method also has the

potential to indicate a candidate region for modification, to increase protein

solubility. All methods are available for free at the ESPRESSO server

(http://mbs.cbrc.jp/ESPRESSO).

© 2013 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

DOI: 10.1002/pmic.201200175

PMID: 23436767 [Indexed for MEDLINE]

1159. BMC Bioinformatics. 2013 Apr 30;14:145. doi: 10.1186/1471-2105-14-145.

Quality control, analysis and secure sharing of Luminex® immunoassay data using

the open source LabKey Server platform.

Eckels J(1), Nathe C, Nelson EK, Shoemaker SG, Nostrand EV, Yates NL, Ashley VC,

Harris LJ, Bollenbeck M, Fong Y, Tomaras GD, Piehler B.

Author information:

(1)LabKey Software, Seattle, WA, USA.

BACKGROUND: Immunoassays that employ multiplexed bead arrays produce high

information content per sample. Such assays are now frequently used to evaluate

humoral responses in clinical trials. Integrated software is needed for the

analysis, quality control, and secure sharing of the high volume of data produced

by such multiplexed assays. Software that facilitates data exchange and provides

flexibility to perform customized analyses (including multiple curve fits and

visualizations of assay performance over time) could increase scientists'

capacity to use these immunoassays to evaluate human clinical trials.

RESULTS: The HIV Vaccine Trials Network and the Statistical Center for HIV/AIDS

Research and Prevention collaborated with LabKey Software to enhance the open

source LabKey Server platform to facilitate workflows for multiplexed bead

assays. This system now supports the management, analysis, quality control, and

secure sharing of data from multiplexed immunoassays that leverage Luminex xMAP®

technology. These assays may be custom or kit-based. Newly added features enable

labs to: (i) import run data from spreadsheets output by Bio-Plex Manager™

software; (ii) customize data processing, curve fits, and algorithms through

scripts written in common languages, such as R; (iii) select script-defined

calculation options through a graphical user interface; (iv) collect custom

metadata for each titration, analyte, run and batch of runs; (v) calculate

dose-response curves for titrations; (vi) interpolate unknown concentrations from

curves for titrated standards; (vii) flag run data for exclusion from analysis;

(viii) track quality control metrics across runs using Levey-Jennings plots; and

(ix) automatically flag outliers based on expected values. Existing system

features allow researchers to analyze, integrate, visualize, export and securely

share their data, as well as to construct custom user interfaces and workflows.

CONCLUSIONS: Unlike other tools tailored for Luminex immunoassays, LabKey Server

allows labs to customize their Luminex analyses using scripting while still

presenting users with a single, graphical interface for processing and analyzing

data. The LabKey Server system also stands out among Luminex tools for enabling

smooth, secure transfer of data, quality control information, and analyses

between collaborators. LabKey Server and its Luminex features are freely

available as open source software at http://www.labkey.com under the Apache 2.0

license.

DOI: 10.1186/1471-2105-14-145

PMCID: PMC3671158

PMID: 23631706 [Indexed for MEDLINE]

1160. J Vis Exp. 2013 Apr 26;(74):e4401. doi: 10.3791/4401.

Analyzing and building nucleic acid structures with 3DNA.

Colasanti AV(1), Lu XJ, Olson WK.

Author information:

(1)Department of Chemistry & Chemical Biology and BioMaPS Institute for

Quantitative Biology, Rutgers - The State University of New Jersey.

andrewco@rci.rutgers.edu

The 3DNA software package is a popular and versatile bioinformatics tool with

capabilities to analyze, construct, and visualize three-dimensional nucleic acid

structures. This article presents detailed protocols for a subset of new and

popular features available in 3DNA, applicable to both individual structures and

ensembles of related structures. Protocol 1 lists the set of instructions needed

to download and install the software. This is followed, in Protocol 2, by the

analysis of a nucleic acid structure, including the assignment of base pairs and

the determination of rigid-body parameters that describe the structure and, in

Protocol 3, by a description of the reconstruction of an atomic model of a

structure from its rigid-body parameters. The most recent version of 3DNA,

version 2.1, has new features for the analysis and manipulation of ensembles of

structures, such as those deduced from nuclear magnetic resonance (NMR)

measurements and molecular dynamic (MD) simulations; these features are presented

in Protocols 4 and 5. In addition to the 3DNA stand-alone software package, the

w3DNA web server, located at http://w3dna.rutgers.edu, provides a user-friendly

interface to selected features of the software. Protocol 6 demonstrates a novel

feature of the site for building models of long DNA molecules decorated with

bound proteins at user-specified locations.

DOI: 10.3791/4401

PMCID: PMC3667640

PMID: 23644419 [Indexed for MEDLINE]

1161. BMC Bioinformatics. 2013 Apr 24;14:140. doi: 10.1186/1471-2105-14-140.

D-Light on promoters: a client-server system for the analysis and visualization

of cis-regulatory elements.

Laimer J(1), Zuzan CJ, Ehrenberger T, Freudenberger M, Gschwandtner S, Lebherz C,

Lackner P.

Author information:

(1)Department of Molecular Biology, University of Salzburg, Hellbrunnerstr, 34,

5020 Salzburg, Austria.

BACKGROUND: The binding of transcription factors to DNA plays an essential role

in the regulation of gene expression. Numerous experiments elucidated binding

sequences which subsequently have been used to derive statistical models for

predicting potential transcription factor binding sites (TFBS). The rapidly

increasing number of genome sequence data requires sophisticated computational

approaches to manage and query experimental and predicted TFBS data in the

context of other epigenetic factors and across different organisms.

RESULTS: We have developed D-Light, a novel client-server software package to

store and query large amounts of TFBS data for any number of genomes. Users can

add small-scale data to the server database and query them in a large scale,

genome-wide promoter context. The client is implemented in Java and provides

simple graphical user interfaces and data visualization. Here we also performed a

statistical analysis showing what a user can expect for certain parameter

settings and we illustrate the usage of D-Light with the help of a microarray

data set.

CONCLUSIONS: D-Light is an easy to use software tool to integrate, store and

query annotation data for promoters. A public D-Light server, the client and

server software for local installation and the source code under GNU GPL license

are available at http://biwww.che.sbg.ac.at/dlight.

DOI: 10.1186/1471-2105-14-140

PMCID: PMC3685601

PMID: 23617301 [Indexed for MEDLINE]

1162. BMC Genomics. 2013 Apr 20;14:269. doi: 10.1186/1471-2164-14-269.

P2RP: a Web-based framework for the identification and analysis of regulatory

proteins in prokaryotic genomes.

Barakat M(1), Ortet P, Whitworth DE.

Author information:

(1)CEA, Saint-Paul-lez-Durance, France. mohamed.barakat@cea.fr

BACKGROUND: Regulatory proteins (RPs) such as transcription factors (TFs) and

two-component system (TCS) proteins control how prokaryotic cells respond to

changes in their external and/or internal state. Identification and annotation of

TFs and TCSs is non-trivial, and between-genome comparisons are often confounded

by different standards in annotation. There is a need for user-friendly, fast and

convenient tools to allow researchers to overcome the inherent variability in

annotation between genome sequences.

RESULTS: We have developed the web-server P2RP (Predicted Prokaryotic Regulatory

Proteins), which enables users to identify and annotate TFs and TCS proteins

within their sequences of interest. Users can input amino acid or genomic DNA

sequences, and predicted proteins therein are scanned for the possession of

DNA-binding domains and/or TCS domains. RPs identified in this manner are

categorised into families, unambiguously annotated, and a detailed description of

their features generated, using an integrated software pipeline. P2RP results can

then be outputted in user-specified formats.

CONCLUSION: Biologists have an increasing need for fast and intuitively usable

tools, which is why P2RP has been developed as an interactive system. As well as

assisting experimental biologists to interrogate novel sequence data, it is hoped

that P2RP will be built into genome annotation pipelines and re-annotation

processes, to increase the consistency of RP annotation in public genomic

sequences. P2RP is the first publicly available tool for predicting and analysing

RP proteins in users' sequences. The server is freely available and can be

accessed along with documentation at http://www.p2rp.org.

DOI: 10.1186/1471-2164-14-269

PMCID: PMC3637814

PMID: 23601859 [Indexed for MEDLINE]

1163. Bioinformatics. 2013 Apr 15;29(8):1078-80. doi: 10.1093/bioinformatics/btt079.

Epub 2013 Feb 14.

GalaxyGemini: a web server for protein homo-oligomer structure prediction based

on similarity.

Lee H(1), Park H, Ko J, Seok C.

Author information:

(1)Department of Chemistry, Seoul National University, Seoul, Republic of Korea.

SUMMARY: A large number of proteins function as homo-oligomers; therefore,

predicting homo-oligomeric structure of proteins is of primary importance for

understanding protein function at the molecular level. Here, we introduce a web

server for prediction of protein homo-oligomer structure. The server takes a

protein monomer structure as input and predicts its homo-oligomer structure from

oligomer templates selected based on sequence and tertiary/quaternary structure

similarity. Using protein model structures as input, the server shows clear

improvement over the best methods of CASP9 in predicting oligomeric structures

from amino acid sequences.

AVAILABILITY: http://galaxy.seoklab.org/gemini.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt079

PMID: 23413437 [Indexed for MEDLINE]

1164. BMC Bioinformatics. 2013 Apr 15;14:130. doi: 10.1186/1471-2105-14-130.

Predicting substrates of the human breast cancer resistance protein using a

support vector machine method.

Hazai E(1), Hazai I, Ragueneau-Majlessi I, Chung SP, Bikadi Z, Mao Q.

Author information:

(1)Virtua Drug Ltd., Csalogany Street 4, Budapest H-1015, Hungary.

BACKGROUND: Human breast cancer resistance protein (BCRP) is an ATP-binding

cassette (ABC) efflux transporter that confers multidrug resistance in cancers

and also plays an important role in the absorption, distribution and elimination

of drugs. Prediction as to if drugs or new molecular entities are BCRP substrates

should afford a cost-effective means that can help evaluate the pharmacokinetic

properties, efficacy, and safety of these drugs or drug candidates. At present,

limited studies have been done to develop in silico prediction models for BCRP

substrates. In this study, we developed support vector machine (SVM) models to

predict wild-type BCRP substrates based on a total of 263 known BCRP substrates

and non-substrates collected from literature. The final SVM model was integrated

to a free web server.

RESULTS: We showed that the final SVM model had an overall prediction accuracy of

~73% for an independent external validation data set of 40 compounds. The

prediction accuracy for wild-type BCRP substrates was ~76%, which is higher than

that for non-substrates. The free web server (http://bcrp.althotas.com) allows

the users to predict whether a query compound is a wild-type BCRP substrate and

calculate its physicochemical properties such as molecular weight, logP value,

and polarizability.

CONCLUSIONS: We have developed an SVM prediction model for wild-type BCRP

substrates based on a relatively large number of known wild-type BCRP substrates

and non-substrates. This model may prove valuable for screening substrates and

non-substrates of BCRP, a clinically important ABC efflux drug transporter.

DOI: 10.1186/1471-2105-14-130

PMCID: PMC3641962

PMID: 23586520 [Indexed for MEDLINE]

1165. BMC Bioinformatics. 2013 Apr 15;14:127. doi: 10.1186/1471-2105-14-127.

VaccImm: simulating peptide vaccination in cancer therapy.

von Eichborn J(1), Woelke AL, Castiglione F, Preissner R.

Author information:

(1)Charité-Universitätsmedizin Berlin, Institute for Physiology, Berlin, Germany.

BACKGROUND: Despite progress in conventional cancer therapies, cancer is still

one of the leading causes of death in industrial nations. Therefore, an urgent

need of progress in fighting cancer remains. A promising alternative to

conventional methods is immune therapy. This relies on the fact that

low-immunogenic tumours can be eradicated if an immune response against them is

induced. Peptide vaccination is carried out by injecting tumour peptides into a

patient to trigger a specific immune response against the tumour in its entirety.

However, peptide vaccination is a highly complicated treatment and currently many

factors like the optimal number of epitopes are not known precisely. Therefore,

it is necessary to evaluate how certain parameters influence the therapy.

RESULTS: We present the VaccImm Server that allows users to simulate peptide

vaccination in cancer therapy. It uses an agent-based model that simulates

peptide vaccination by explicitly modelling the involved cells (immune system and

cancer) as well as molecules (antibodies, antigens and semiochemicals). As a new

feature, our model uses real amino acid sequences to represent molecular binding

sites of relevant immune cells. The model is used to generate detailed statistics

of the population sizes and states of the single cell types over time. This makes

the VaccImm web server well suited to examine the parameter space of peptide

vaccination in silico. VaccImm is publicly available without registration on the

web at http://bioinformatics.charite.de/vaccimm; all major browsers are

supported.

CONCLUSIONS: The VaccImm Server provides a convenient way to analyze properties

of peptide vaccination in cancer therapy. Using the server, we could gain

interesting insights into peptide vaccination that reveal the complex and

patient-specific nature of peptide vaccination.

DOI: 10.1186/1471-2105-14-127

PMCID: PMC3651379

PMID: 23586423 [Indexed for MEDLINE]

1166. Gene. 2013 Apr 10;518(1):114-23. doi: 10.1016/j.gene.2012.11.063. Epub 2013 Jan

12.

GI-POP: a combinational annotation and genomic island prediction pipeline for

ongoing microbial genome projects.

Lee CC(1), Chen YP, Yao TJ, Ma CY, Lo WC, Lyu PC, Tang CY.

Author information:

(1)Institute of Bioinformatics and Structural Biology, National Tsing Hua

University, Hsinchu, Taiwan. d9662841@oz.nthu.edu.tw

Sequencing of microbial genomes is important because of microbial-carrying

antibiotic and pathogenetic activities. However, even with the help of new

assembling software, finishing a whole genome is a time-consuming task. In most

bacteria, pathogenetic or antibiotic genes are carried in genomic islands.

Therefore, a quick genomic island (GI) prediction method is useful for ongoing

sequencing genomes. In this work, we built a Web server called GI-POP

(http://gipop.life.nthu.edu.tw) which integrates a sequence assembling tool, a

functional annotation pipeline, and a high-performance GI predicting module, in a

support vector machine (SVM)-based method called genomic island genomic profile

scanning (GI-GPS). The draft genomes of the ongoing genome projects in contigs or

scaffolds can be submitted to our Web server, and it provides the functional

annotation and highly probable GI-predicting results. GI-POP is a comprehensive

annotation Web server designed for ongoing genome project analysis. Researchers

can perform annotation and obtain pre-analytic information include possible GIs,

coding/non-coding sequences and functional analysis from their draft genomes.

This pre-analytic system can provide useful information for finishing a genome

sequencing project.

Copyright © 2012 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.gene.2012.11.063

PMID: 23318308 [Indexed for MEDLINE]

1167. Gene. 2013 Apr 10;518(1):78-83. doi: 10.1016/j.gene.2012.11.083. Epub 2012 Dec

28.

PRASA: an integrated web server that analyzes protein interaction types.

Fan CY(1), Bai YH, Huang CY, Yao TJ, Chiang WH, Chang DT.

Author information:

(1)Department of Electrical Engineering, National Cheng Kung University, Tainan

70101, Taiwan.

This work presents the Protein Association Analyzer (PRASA)

(http://zoro.ee.ncku.edu.tw/prasa/) that predicts protein interactions as well as

interaction types. Protein interactions are essential to most biological

functions. The existence of diverse interaction types, such as physically

contacted or functionally related interactions, makes protein interactions

complex. Different interaction types are distinct and should not be confused.

However, most existing tools focus on a specific interaction type or mix

different interaction types. This work collected 7234058 associations with

experimentally verified interaction types from five databases and compiled

individual probabilistic models for different interaction types. The PRASA result

page shows predicted associations and their related references by interaction

type. Experimental results demonstrate the performance difference when

distinguishing between different interaction types. The PRASA provides a

centralized and organized platform for easy browsing, downloading and comparing

of interaction types, which helps reveal insights into the complex roles that

proteins play in organisms.

Copyright © 2012 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.gene.2012.11.083

PMID: 23276706 [Indexed for MEDLINE]

1168. Gene. 2013 Apr 10;518(1):26-34. doi: 10.1016/j.gene.2012.11.089. Epub 2012 Dec

22.

YGA: identifying distinct biological features between yeast gene sets.

Chang DT(1), Li WS, Bai YH, Wu WS.

Author information:

(1)Department of Electrical Engineering, National Cheng Kung University, Tainan

70101, Taiwan.

The advance of high-throughput experimental technologies generates many gene sets

with different biological meanings, where many important insights can only be

extracted by identifying the biological (regulatory/functional) features that are

distinct between different gene sets (e.g. essential vs. non-essential genes,

TATA box-containing vs. TATA box-less genes, induced vs. repressed genes under

certain biological conditions). Although many servers have been developed to

identify enriched features in a gene set, most of them were designed to analyze

one gene set at a time but cannot compare two gene sets. Moreover, the features

used in existing servers were mainly focused on functional annotations (GO

terms), pathways, transcription factor binding sites (TFBSs) and/or

protein-protein interactions (PPIs). In yeast, various important regulatory

features, including promoter bendability, nucleosome occupancy, 5'-UTR length,

and TF-gene regulation evidence, are available but have not been used in any

enrichment analysis servers. This motivates us to develop the Yeast Genes

Analyzer (YGA), a web server that simultaneously analyzes various biological

(regulatory/functional) features of two gene sets and performs statistical tests

to identify the distinct features between them. Many well-studied gene sets such

as essential, stress-response, TATA box-containing and cell cycle genes were

pre-compiled in YGA for users, if they have only one gene set, to compare with.

In comparison with the existing enrichment analysis servers, YGA tests more

comprehensive regulatory features (e.g. promoter bendability, nucleosome

occupancy, 5'-UTR length, experimental evidence of TF-gene binding and TF-gene

regulation) and functional features (e.g. PPI, GO terms, pathways and functional

groups of genes, including essential/non-essential genes,

stress-induced/-repressed genes, TATA box-containing/-less genes,

occupied/depleted proximal-nucleosome genes and cell cycle genes). Furthermore,

YGA uses various statistical tests to provide objective comparison measures. The

two major contributions of YGA, comprehensive features and statistical

comparison, help to mine important information that cannot be obtained from other

servers. The sophisticated analysis tools of YGA can identify distinct biological

features between two gene sets, which help biologists to form new hypotheses

about the underlying biological mechanisms responsible for the observed

difference between these two gene sets. YGA can be accessed from the following

web pages: http://cosbi.ee.ncku.edu.tw/yga/ and http://yga.ee.ncku.edu.tw/.

Copyright © 2012 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.gene.2012.11.089

PMID: 23266802 [Indexed for MEDLINE]

1169. J Proteome Res. 2013 Apr 5;12(4):1989-95. doi: 10.1021/pr301162j. Epub 2013 Mar

5.

XLink-DB: database and software tools for storing and visualizing protein

interaction topology data.

Zheng C(1), Weisbrod CR, Chavez JD, Eng JK, Sharma V, Wu X, Bruce JE.

Author information:

(1)Department of Chemistry, University of Washington , Seattle, Washington,

United States.

As large-scale cross-linking data becomes available, new software tools for data

processing and visualization are required to replace manual data analysis.

XLink-DB serves as a data storage site and visualization tool for cross-linking

results. XLink-DB accepts data generated with any cross-linker and stores them in

a relational database. Cross-linked sites are automatically mapped onto PDB

structures if available, and results are compared to existing protein interaction

databases. A protein interaction network is also automatically generated for the

entire data set. The XLink-DB server, including examples, and a help page are

available for noncommercial use at

http://brucelab.gs.washington.edu/crosslinkdbv1/ . The source code can be viewed

and downloaded at https://sourceforge.net/projects/crosslinkdb/?source=directory

.

DOI: 10.1021/pr301162j

PMCID: PMC3744611

PMID: 23413830 [Indexed for MEDLINE]

1170. Bioinformatics. 2013 Apr 1;29(7):925-32. doi: 10.1093/bioinformatics/btt061. Epub

2013 Feb 15.

Chemical rule-based filtering of MS/MS spectra.

Reiz B(1), Kertész-Farkas A, Pongor S, Myers MP.

Author information:

(1)Protein Structure and Bioinformatics Group, International Centre for Genetic

Engineering and Biotechnology, 34149 Trieste, Italy.

MOTIVATION: Identification of proteins by mass spectrometry-based proteomics

requires automated interpretation of peptide tandem mass spectrometry spectra.

The effectiveness of peptide identification can be greatly improved by filtering

out extraneous noise peaks before the subsequent database searching steps.

RESULTS: Here we present a novel chemical rule-based filtering algorithm, termed

CRF, which makes use of the predictable patterns (rules) of collision-induced

peptide fragmentation. The algorithm selects peak pairs that obey the common

fragmentation rules within plausible limits of mass tolerance as well as peak

intensity and produces spectra that can be subsequently submitted to any search

engine. CRF increases the positive predictive value and decreases the number of

random matches and thus improves performance by 15-20% in terms of peptide

annotation using search engines, such as X!Tandem. Importantly, the algorithm

also achieves data compression rates of ∼75%.

AVAILABILITY: The MATLAB source code and a web server are available at

http://hydrax.icgeb.trieste.it/CRFilter/.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btt061

PMID: 23418188 [Indexed for MEDLINE]

1171. Bioinformatics. 2013 Apr 1;29(7):947-9. doi: 10.1093/bioinformatics/btt064. Epub

2013 Feb 8.

Phylogenomic clustering for selecting non-redundant genomes for comparative

genomics.

Moreno-Hagelsieb G(1), Wang Z, Walsh S, ElSherbiny A.

Author information:

(1)Department of Biology and Department of Mathematics, Wilfrid Laurier

University, Waterloo, ON N2L 3C5, Canada. gmoreno@wlu.ca

MOTIVATION: Analyses in comparative genomics often require non-redundant genome

datasets. Eliminating redundancy is not as simple as keeping one strain for each

named species because genomes might be redundant at a higher taxonomic level than

that of species for some analyses; some strains with different species names can

be as similar as most strains sharing a species name, whereas some strains

sharing a species name can be so different that they should be put into different

groups; and some genomes lack a species name.

RESULTS: We have implemented a method and Web server that clusters a genome

dataset into groups of redundant genomes at different thresholds based on a few

phylogenomic distance measures.

AVAILABILITY: The Web interface, similarity and distance data and R-scripts can

be accessed at http://microbiome.wlu.ca/research/redundancy/.

DOI: 10.1093/bioinformatics/btt064

PMID: 23396122 [Indexed for MEDLINE]

1172. Hum Mutat. 2013 Apr;34(4):557-65. doi: 10.1002/humu.22277. Epub 2013 Feb 21.

Prediction of mutant mRNA splice isoforms by information theory-based exon

definition.

Mucaki EJ(1), Shirley BC, Rogan PK.

Author information:

(1)Department of Biochemistry, Western University, London, Ontario, Canada.

Mutations that affect mRNA splicing often produce multiple mRNA isoforms,

resulting in complex molecular phenotypes. Definition of an exon and its

inclusion in mature mRNA relies on joint recognition of both acceptor and donor

splice sites. This study predicts cryptic and exon-skipping isoforms in mRNA

produced by splicing mutations from the combined information contents (R(i),

which measures binding-site strength, in bits) and distribution of the splice

sites defining these exons. The total information content of an exon (R(i),total)

is the sum of the R(i) values of its acceptor and donor splice sites, adjusted

for the self-information of the distance separating these sites, that is, the gap

surprisal. Differences between total information contents of an exon

(ΔR(i,total)) are predictive of the relative abundance of these exons in distinct

processed mRNAs. Constraints on splice site and exon selection are used to

eliminate nonconforming and poorly expressed isoforms. Molecular phenotypes are

computed by the Automated Splice Site and Exon Definition Analysis

(http://splice.uwo.ca) server. Predictions of splicing mutations were highly

concordant (85.2%; n = 61) with published expression data. In silico exon

definition analysis will contribute to streamlining assessment of abnormal and

normal splice isoforms resulting from mutations.

© 2013 Wiley Periodicals, Inc.

DOI: 10.1002/humu.22277

PMID: 23348723 [Indexed for MEDLINE]

1173. J Lab Autom. 2013 Apr;18(2):137-42. doi: 10.1177/2211068212471671. Epub 2012 Dec

27.

Development of advanced fermentor control applications for use in an industrial

automation environment.

Hamilton R(1), Tamminana K, Boyd J, Sasaki G, Toda A, Haskell S, Danbe E.

Author information:

(1)Pharma Technical Development, Genentech, South San Francisco, CA 94080, USA.

hamilton.ryan@gene.com

We present a software platform developed by Genentech and MathWorks Consulting

Group that allows arbitrary MATLAB (MATLAB is a registered trademark of The

MathWorks, Inc.) functions to perform supervisory control of process equipment

(in this case, fermentors) via the OLE for process control (OPC) communication

protocol, under the direction of an industrial automation layer. The software

features automated synchronization and deployment of server control code and has

been proven to be tolerant of OPC communication interruptions. Since deployment

in the spring of 2010, this software has successfully performed supervisory

control of more than 700 microbial fermentations in the Genentech pilot plant and

has enabled significant reductions in the time required to develop and implement

novel control strategies (months reduced to days). The software is available for

download at the MathWorks File Exchange Web site at

http://www.mathworks.com/matlabcentral/fileexchange/36866.

DOI: 10.1177/2211068212471671

PMID: 23271785 [Indexed for MEDLINE]

1174. Nucleic Acids Res. 2013 Apr 1;41(6):e68. doi: 10.1093/nar/gks1450. Epub 2013 Jan

8.

iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition.

Chen W(1), Feng PM, Lin H, Chou KC.

Author information:

(1)Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, Hebei United University, Tangshan, China.

wchen@gordonlifescience.org

Meiotic recombination is an important biological process. As a main driving force

of evolution, recombination provides natural new combinations of genetic

variations. Rather than randomly occurring across a genome, meiotic recombination

takes place in some genomic regions (the so-called 'hotspots') with higher

frequencies, and in the other regions (the so-called 'coldspots') with lower

frequencies. Therefore, the information of the hotspots and coldspots would

provide useful insights for in-depth studying of the mechanism of recombination

and the genome evolution process as well. So far, the recombination regions have

been mainly determined by experiments, which are both expensive and

time-consuming. With the avalanche of genome sequences generated in the

postgenomic age, it is highly desired to develop automated methods for rapidly

and effectively identifying the recombination regions. In this study, a

predictor, called 'iRSpot-PseDNC', was developed for identifying the

recombination hotspots and coldspots. In the new predictor, the samples of DNA

sequences are formulated by a novel feature vector, the so-called 'pseudo

dinucleotide composition' (PseDNC), into which six local DNA structural

properties, i.e. three angular parameters (twist, tilt and roll) and three

translational parameters (shift, slide and rise), are incorporated. It was

observed by the rigorous jackknife test that the overall success rate achieved by

iRSpot-PseDNC was >82% in identifying recombination spots in Saccharomyces

cerevisiae, indicating the new predictor is promising or at least may become a

complementary tool to the existing methods in this area. Although the benchmark

data set used to train and test the current method was from S. cerevisiae, the

basic approaches can also be extended to deal with all the other genomes.

Particularly, it has not escaped our notice that the PseDNC approach can be also

used to study many other DNA-related problems. As a user-friendly web-server,

iRSpot-PseDNC is freely accessible at

http://lin.uestc.edu.cn/server/iRSpot-PseDNC.

DOI: 10.1093/nar/gks1450

PMCID: PMC3616736

PMID: 23303794 [Indexed for MEDLINE]

1175. Plant J. 2013 Apr;74(2):351-62. doi: 10.1111/tpj.12119. Epub 2013 Mar 4.

Establishment of the Lotus japonicus Gene Expression Atlas (LjGEA) and its use to

explore legume seed maturation.

Verdier J(1), Torres-Jerez I, Wang M, Andriankaja A, Allen SN, He J, Tang Y,

Murray JD, Udvardi MK.

Author information:

(1)Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401,

USA.

Lotus japonicus is a model species for legume genomics. To accelerate legume

functional genomics, we developed a Lotus japonicus Gene Expression Atlas

(LjGEA), which provides a global view of gene expression in all organ systems of

this species, including roots, nodules, stems, petioles, leaves, flowers, pods

and seeds. Time-series data covering multiple stages of developing pod and seed

are included in the LjGEA. In addition, previously published L. japonicus

Affymetrix data are included in the database, making it a 'one-stop shop' for

transcriptome analysis of this species. The LjGEA web server

(http://ljgea.noble.org/) enables flexible, multi-faceted analyses of the

transcriptome. Transcript data may be accessed using the Affymetrix probe

identification number, DNA sequence, gene name, functional description in natural

language, and GO and KEGG annotation terms. Genes may be discovered through

co-expression or differential expression analysis. Users may select a subset of

experiments and visualize and compare expression profiles of multiple genes

simultaneously. Data may be downloaded in a tabular form compatible with common

analytical and visualization software. To illustrate the power of LjGEA, we

explored the transcriptome of developing seeds. Genes represented by 36 474 probe

sets were expressed at some stage during seed development, and almost half of

these genes displayed differential expression during development. Among the

latter were 624 transcription factor genes, some of which are orthologs of

transcription factor genes that are known to regulate seed development in other

species, while most are novel and represent attractive targets for reverse

genetics approaches to determine their roles in this important organ.

© 2013 The Authors The Plant Journal © 2013 Blackwell Publishing Ltd.

DOI: 10.1111/tpj.12119

PMID: 23452239 [Indexed for MEDLINE]

1176. Protein Eng Des Sel. 2013 Apr;26(4):283-9. doi: 10.1093/protein/gzs108. Epub 2013

Jan 21.

A structural bioinformatics approach for identifying proteins predisposed to bind

linear epitopes on pre-selected target proteins.

Choi EJ(1), Jacak R, Kuhlman B.

Author information:

(1)Department of Biochemistry and Biophysics, University of North Carolina,

Chapel Hill, NC 27599, USA.

We have developed a protocol for identifying proteins that are predisposed to

bind linear epitopes on target proteins of interest. The protocol searches

through the protein database for proteins (scaffolds) that are bound to peptides

with sequences similar to accessible, linear epitopes on the target protein. The

sequence match is considered more significant if residues calculated to be

important in the scaffold-peptide interaction are present in the target epitope.

The crystal structure of the scaffold-peptide complex is then used as a template

for creating a model of the scaffold bound to the target epitope. This model can

then be used in conjunction with sequence optimization algorithms or directed

evolution methods to search for scaffold mutations that further increase affinity

for the target protein. To test the applicability of this approach we targeted

three disease-causing proteins: a tuberculosis virulence factor (TVF), the apical

membrane antigen (AMA) from malaria, and hemagglutinin from influenza. In each

case the best scoring scaffold was tested, and binders with Kds equal to 37 μM

and 50 nM for TVF and AMA, respectively, were identified. A web server

(http://rosettadesign.med.unc.edu/scaffold/) has been created for performing the

scaffold search process with user-defined target sequences.

DOI: 10.1093/protein/gzs108

PMCID: PMC3601849

PMID: 23341643 [Indexed for MEDLINE]

1177. BMC Bioinformatics. 2013 Mar 27;14:111. doi: 10.1186/1471-2105-14-111.

A benchmark server using high resolution protein structure data, and benchmark

results for membrane helix predictions.

Rath EM(1), Tessier D, Campbell AA, Lee HC, Werner T, Salam NK, Lee LK, Church

WB.

Author information:

(1)Group in Biomolecular Structure and Informatics, Faculty of Pharmacy, The

University of Sydney, Darlinghurst, Sydney NSW 2006, Australia.

BACKGROUND: Helical membrane proteins are vital for the interaction of cells with

their environment. Predicting the location of membrane helices in protein amino

acid sequences provides substantial understanding of their structure and function

and identifies membrane proteins in sequenced genomes. Currently there is no

comprehensive benchmark tool for evaluating prediction methods, and there is no

publication comparing all available prediction tools. Current benchmark

literature is outdated, as recently determined membrane protein structures are

not included. Current literature is also limited to global assessments, as

specialised benchmarks for predicting specific classes of membrane proteins were

not previously carried out.

DESCRIPTION: We present a benchmark server at

http://sydney.edu.au/pharmacy/sbio/software/TMH\_benchmark.shtml that uses recent

high resolution protein structural data to provide a comprehensive assessment of

the accuracy of existing membrane helix prediction methods. The server further

allows a user to compare uploaded predictions generated by novel methods,

permitting the comparison of these novel methods against all existing methods

compared by the server. Benchmark metrics include sensitivity and specificity of

predictions for membrane helix location and orientation, and many others. The

server allows for customised evaluations such as assessing prediction method

performances for specific helical membrane protein subtypes.We report results for

custom benchmarks which illustrate how the server may be used for specialised

benchmarks. Which prediction method is the best performing method depends on

which measure is being benchmarked. The OCTOPUS membrane helix prediction method

is consistently one of the highest performing methods across all measures in the

benchmarks that we performed.

CONCLUSIONS: The benchmark server allows general and specialised assessment of

existing and novel membrane helix prediction methods. Users can employ this

benchmark server to determine the most suitable method for the type of prediction

the user needs to perform, be it general whole-genome annotation or the

prediction of specific types of helical membrane protein. Creators of novel

prediction methods can use this benchmark server to evaluate the performance of

their new methods. The benchmark server will be a valuable tool for researchers

seeking to extract more sophisticated information from the large and growing

protein sequence databases.

DOI: 10.1186/1471-2105-14-111

PMCID: PMC3620685

PMID: 23530628 [Indexed for MEDLINE]

1178. Chem Biol Interact. 2013 Mar 25;203(1):266-8. doi: 10.1016/j.cbi.2012.09.003.

Epub 2012 Sep 23.

Proteins with an alpha/beta hydrolase fold: Relationships between subfamilies in

an ever-growing superfamily.

Lenfant N(1), Hotelier T, Bourne Y, Marchot P, Chatonnet A.

Author information:

(1)Dynamique Musculaire et Métabolisme, INRA, Place Viala, Montpellier, France.

Alpha/beta hydrolases function as hydrolases, lyases, transferases, hormone

precursors or transporters, chaperones or routers of other proteins. The amount

of structural and functional available data related to this protein superfamily

expands exponentially, as does the number of proteins classified as alpha/beta

hydrolases despite poor sequence similarity and lack of experimental data.

However the superfamily can be rationally divided according to sequence or

structural homologies, leading to subfamilies of proteins with potentially

similar functions. Since the discovery of proteins homologous to cholinesterases

but devoid of enzymatic activity (e.g., the neuroligins), divergent functions

have been ascribed to members of other subfamilies (e.g., lipases,

dipeptidylaminopeptidase IV, etc.). To study the potentially moonlighting

properties of alpha/beta hydrolases, the ESTHER database (for ESTerase and

alpha/beta Hydrolase Enzymes and Relatives; http://bioweb.ensam.inra.fr/esther),

which collects, organizes and disseminates structural and functional information

related to alpha/beta hydrolases, has been updated with new tools and the web

server interface has been upgraded. A new Overall Table along with a new Tree

based on HMM models has been included to tentatively group subfamilies. These

tools provide starting points for phylogenetic studies aimed at pinpointing the

origin of duplications leading to paralogous genes (e.g., acetylcholinesterase

versus butyrylcholinesterase, or neuroligin versus carboxylesterase). Another of

our goals is to implement new tools to distinguish catalytically active enzymes

from non-catalytic proteins in poorly studied or annotated subfamilies.

Copyright © 2012 Elsevier Ireland Ltd. All rights reserved.

DOI: 10.1016/j.cbi.2012.09.003

PMID: 23010363 [Indexed for MEDLINE]

1179. J Transl Med. 2013 Mar 22;11:74. doi: 10.1186/1479-5876-11-74.

In silico approaches for designing highly effective cell penetrating peptides.

Gautam A(1), Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A; Open source drug

discovery consortium, Raghava GP.

Author information:

(1)Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh

160036, India.

BACKGROUND: Cell penetrating peptides have gained much recognition as a versatile

transport vehicle for the intracellular delivery of wide range of cargoes (i.e.

oligonucelotides, small molecules, proteins, etc.), that otherwise lack

bioavailability, thus offering great potential as future therapeutics. Keeping in

mind the therapeutic importance of these peptides, we have developed in silico

methods for the prediction of cell penetrating peptides, which can be used for

rapid screening of such peptides prior to their synthesis.

METHODS: In the present study, support vector machine (SVM)-based models have

been developed for predicting and designing highly effective cell penetrating

peptides. Various features like amino acid composition, dipeptide composition,

binary profile of patterns, and physicochemical properties have been used as

input features. The main dataset used in this study consists of 708 peptides. In

addition, we have identified various motifs in cell penetrating peptides, and

used these motifs for developing a hybrid prediction model. Performance of our

method was evaluated on an independent dataset and also compared with that of the

existing methods.

RESULTS: In cell penetrating peptides, certain residues (e.g. Arg, Lys, Pro, Trp,

Leu, and Ala) are preferred at specific locations. Thus, it was possible to

discriminate cell-penetrating peptides from non-cell penetrating peptides based

on amino acid composition. All models were evaluated using five-fold

cross-validation technique. We have achieved a maximum accuracy of 97.40% using

the hybrid model that combines motif information and binary profile of the

peptides. On independent dataset, we achieved maximum accuracy of 81.31% with MCC

of 0.63.

CONCLUSION: The present study demonstrates that features like amino acid

composition, binary profile of patterns and motifs, can be used to train an SVM

classifier that can predict cell penetrating peptides with higher accuracy. The

hybrid model described in this study achieved more accuracy than the previous

methods and thus may complement the existing methods. Based on the above study, a

user-friendly web server CellPPD has been developed to help the biologists, where

a user can predict and design CPPs with much ease. CellPPD web server is freely

accessible at http://crdd.osdd.net/raghava/cellppd/.

DOI: 10.1186/1479-5876-11-74

PMCID: PMC3615965

PMID: 23517638 [Indexed for MEDLINE]

1180. Int J Mol Sci. 2013 Mar 18;14(3):6144-56. doi: 10.3390/ijms14036144.

CentroidAlign-Web: A Fast and Accurate Multiple Aligner for Long Non-Coding RNAs.

Yonemoto H(1), Asai K, Hamada M.

Author information:

(1)Department of Computational Biology, Graduate School of Frontier Sciences,the

University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8561, Japan.

mhamada@k.u-tokyo.ac.jp.

Due to the recent discovery of non-coding RNAs (ncRNAs), multiple sequence

alignment (MSA) of those long RNA sequences is becoming increasingly important

for classifying and determining the functional motifs in RNAs. However, not only

primary (nucleotide) sequences, but also secondary structures of ncRNAs are

closely related to their function and are conserved evolutionarily. Hence,

information about secondary structures should be considered in the sequence

alignment of ncRNAs. Yet, in general, a huge computational time is required in

order to compute MSAs, taking secondary structure information into account. In

this paper, we describe a fast and accurate web server, called CentroidAlign-Web,

which can handle long RNA sequences. The web server also appropriately

incorporates information about known secondary structures into MSAs.

Computational experiments indicate that our web server is fast and accurate

enough to handle long RNA sequences. CentroidAlign-Web is freely available from

http://centroidalign.ncrna.org/.

DOI: 10.3390/ijms14036144

PMCID: PMC3634467

PMID: 23507751

1181. BMC Plant Biol. 2013 Mar 15;13:42. doi: 10.1186/1471-2229-13-42.

Promzea: a pipeline for discovery of co-regulatory motifs in maize and other

plant species and its application to the anthocyanin and phlobaphene biosynthetic

pathways and the Maize Development Atlas.

Liseron-Monfils C(1), Lewis T, Ashlock D, McNicholas PD, Fauteux F, Strömvik M,

Raizada MN.

Author information:

(1)Department of Plant Agriculture, University of Guelph, Guelph, Ontario,

Canada.

BACKGROUND: The discovery of genetic networks and cis-acting DNA motifs

underlying their regulation is a major objective of transcriptome studies. The

recent release of the maize genome (Zea mays L.) has facilitated in silico

searches for regulatory motifs. Several algorithms exist to predict cis-acting

elements, but none have been adapted for maize.

RESULTS: A benchmark data set was used to evaluate the accuracy of three motif

discovery programs: BioProspector, Weeder and MEME. Analysis showed that each

motif discovery tool had limited accuracy and appeared to retrieve a distinct set

of motifs. Therefore, using the benchmark, statistical filters were optimized to

reduce the false discovery ratio, and then remaining motifs from all programs

were combined to improve motif prediction. These principles were integrated into

a user-friendly pipeline for motif discovery in maize called Promzea, available

at http://www.promzea.org and on the Discovery Environment of the iPlant

Collaborative website. Promzea was subsequently expanded to include rice and

Arabidopsis. Within Promzea, a user enters cDNA sequences or gene IDs;

corresponding upstream sequences are retrieved from the maize genome. Predicted

motifs are filtered, combined and ranked. Promzea searches the chosen plant

genome for genes containing each candidate motif, providing the user with the

gene list and corresponding gene annotations. Promzea was validated in silico

using a benchmark data set: the Promzea pipeline showed a 22% increase in

nucleotide sensitivity compared to the best standalone program tool, Weeder, with

equivalent nucleotide specificity. Promzea was also validated by its ability to

retrieve the experimentally defined binding sites of transcription factors that

regulate the maize anthocyanin and phlobaphene biosynthetic pathways. Promzea

predicted additional promoter motifs, and genome-wide motif searches by Promzea

identified 127 non-anthocyanin/phlobaphene genes that each contained all five

predicted promoter motifs in their promoters, perhaps uncovering a broader

co-regulated gene network. Promzea was also tested against tissue-specific

microarray data from maize.

CONCLUSIONS: An online tool customized for promoter motif discovery in plants has

been generated called Promzea. Promzea was validated in silico by its ability to

retrieve benchmark motifs and experimentally defined motifs and was tested using

tissue-specific microarray data. Promzea predicted broader networks of gene

regulation associated with the historic anthocyanin and phlobaphene biosynthetic

pathways. Promzea is a new bioinformatics tool for understanding transcriptional

gene regulation in maize and has been expanded to include rice and Arabidopsis.

DOI: 10.1186/1471-2229-13-42

PMCID: PMC3658923

PMID: 23497159 [Indexed for MEDLINE]

1182. J Comput Chem. 2013 Mar 15;34(7):566-75. doi: 10.1002/jcc.23168. Epub 2012 Nov 1.

An accurate and efficient method to predict the electronic excitation energies of

BODIPY fluorescent dyes.

Wang JN(1), Jin JL, Geng Y, Sun SL, Xu HL, Lu YH, Su ZM.

Author information:

(1)Institute of Functional Material Chemistry, Faculty of Chemistry, Northeast

Normal University, Changchun 130024, Jilin, People's Republic of China.

Recently, the extreme learning machine neural network (ELMNN) as a valid

computing method has been proposed to predict the nonlinear optical property

successfully (Wang et al., J. Comput. Chem. 2012, 33, 231). In this work, first,

we follow this line of work to predict the electronic excitation energies using

the ELMNN method. Significantly, the root mean square deviation of the predicted

electronic excitation energies of 90 4,4-difluoro-4-bora-3a,4a-diaza-s-indacene

(BODIPY) derivatives between the predicted and experimental values has been

reduced to 0.13 eV. Second, four groups of molecule descriptors are considered

when building the computing models. The results show that the quantum chemical

descriptions have the closest intrinsic relation with the electronic excitation

energy values. Finally, a user-friendly web server (EEEBPre: Prediction of

electronic excitation energies for BODIPY dyes), which is freely accessible to

public at the web site: http://202.198.129.218, has been built for prediction.

This web server can return the predicted electronic excitation energy values of

BODIPY dyes that are high consistent with the experimental values. We hope that

this web server would be helpful to theoretical and experimental chemists in

related research.

Copyright © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23168

PMID: 23115129 [Indexed for MEDLINE]

1183. IEEE/ACM Trans Comput Biol Bioinform. 2013 Mar 12. [Epub ahead of print]

Multi-Label Learning via Random Label Selection for Protein Subcellular

Multi-Locations Prediction.

Wang X(1), Li GZ.

Author information:

(1)Tongji University, Shanghai and Zhengzhou University of Light Industry,

Zhengzhou.

Prediction of protein subcellular localization is an important but challenging

problem, particularly when proteins may simultaneously exist at, or move between,

two or more different subcellular location sites. Most of the existing protein

subcellular localization methods are only used to deal with the single-location

proteins. In the past few years, only a few methods have been proposed to tackle

proteins with multiple locations. However, they only adopt a simple strategy,

that is, transforming the multi-location proteins to multiple proteins with

single location, which doesn't take correlations among different subcellular

locations into account. In this paper, a novel method named RALS (multi-label

learning via RAndom Label Selection), is proposed to learn from multi-location

proteins in an effective and efficient way. Through five-fold cross validation

test on a benchmark dataset, we demonstrate our proposed method with

consideration of label correlations obviously outperforms the baseline BR method

without consideration of label correlations, indicating correlations among

different subcellular locations really exist and contribute to improvement of

prediction performance. Experimental results on two benchmark datasets also show

that our proposed methods achieve significantly higher performance than some

other state-of-the-art methods in predicting subcellular multi-locations of

proteins. The prediction web server is available at

http://levis.tongji.edu.cn:8080/bioinfo/MLPred-Euk/ for the public usage.

DOI: 2A882EE2-2EC1-435F-B1ED-1CF126E41834

PMID: 23509190

1184. Source Code Biol Med. 2013 Mar 11;8(1):8. doi: 10.1186/1751-0473-8-8.

SVAw - a web-based application tool for automated surrogate variable analysis of

gene expression studies.

Pirooznia M(1), Seifuddin F, Goes FS, Leek JT, Zandi PP.

Author information:

(1)Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of

Medicine, Baltimore, MD, USA. pzandi@jhsph.edu.

BACKGROUND: Surrogate variable analysis (SVA) is a powerful method to identify,

estimate, and utilize the components of gene expression heterogeneity due to

unknown and/or unmeasured technical, genetic, environmental, or demographic

factors. These sources of heterogeneity are common in gene expression studies,

and failing to incorporate them into the analysis can obscure results. Using SVA

increases the biological accuracy and reproducibility of gene expression studies

by identifying these sources of heterogeneity and correctly accounting for them

in the analysis.

RESULTS: Here we have developed a web application called SVAw (Surrogate variable

analysis Web app) that provides a user friendly interface for SVA analyses of

genome-wide expression studies. The software has been developed based on open

source bioconductor SVA package. In our software, we have extended the SVA

program functionality in three aspects: (i) the SVAw performs a fully automated

and user friendly analysis workflow; (ii) It calculates probe/gene Statistics for

both pre and post SVA analysis and provides a table of results for the regression

of gene expression on the primary variable of interest before and after

correcting for surrogate variables; and (iii) it generates a comprehensive report

file, including graphical comparison of the outcome for the user.

CONCLUSIONS: SVAw is a web server freely accessible solution for the surrogate

variant analysis of high-throughput datasets and facilitates removing all

unwanted and unknown sources of variation. It is freely available for use at

http://psychiatry.igm.jhmi.edu/sva. The executable packages for both web and

standalone application and the instruction for installation can be downloaded

from our web site.

DOI: 10.1186/1751-0473-8-8

PMCID: PMC3614430

PMID: 23497726

1185. Structure. 2013 Mar 5;21(3):321-31. doi: 10.1016/j.str.2013.02.004.

All-atom ensemble modeling to analyze small-angle x-ray scattering of

glycosylated proteins.

Guttman M(1), Weinkam P, Sali A, Lee KK.

Author information:

(1)Department of Medicinal Chemistry, University of Washington, Seattle, WA

98195, USA.

The flexible and heterogeneous nature of carbohydrate chains often renders

glycoproteins refractory to traditional structure determination methods.

Small-angle X-ray scattering (SAXS) can be a useful tool for obtaining structural

information of these systems. All-atom modeling of glycoproteins with flexible

glycan chains was applied to interpret the solution SAXS data for a set of

glycoproteins. For simpler systems (single glycan, with a well-defined protein

structure), all-atom modeling generates models in excellent agreement with the

scattering pattern and reveals the approximate spatial occupancy of the glycan

chain in solution. For more complex systems (several glycan chains, or unknown

protein substructure), the approach can still provide insightful models, though

the orientations of glycans become poorly determined. Ab initio shape

reconstructions appear to capture the global morphology of glycoproteins but in

most cases offer little information about glycan spatial occupancy. The all-atom

modeling methodology is available as a web server at

http://salilab.org/allosmod-foxs.

Copyright © 2013 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.str.2013.02.004

PMCID: PMC3840220

PMID: 23473666 [Indexed for MEDLINE]

1186. F1000Res. 2013 Mar 4;2:68. doi: 10.12688/f1000research.2-68.v1. eCollection 2013.

The Focal Adhesion Analysis Server: a web tool for analyzing focal adhesion

dynamics.

Berginski ME(1), Gomez SM(2).

Author information:

(1)UNC/NCSU Joint Department of Biomedical Engineering, University of North

Carolina at Chapel Hill, Chapel Hill, NC, 27599-7575, USA. (2)UNC/NCSU Joint

Department of Biomedical Engineering, University of North Carolina at Chapel

Hill, Chapel Hill, NC, 27599-7575, USA ; UNC Department of Pharmacology,

University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-7575, USA ;

UNC Department of Computer Science, University of North Carolina at Chapel Hill,

Chapel Hill, NC, 27599-7575, USA.

The Focal Adhesion Analysis Server (FAAS) is a web-based implementation of a set

of computer vision algorithms designed to quantify the behavior of focal

adhesions in cells imaged in 2D cultures. The input consists of one or more

images of a labeled focal adhesion protein. The outputs of the system include a

range of static and dynamic measurements for the adhesions present in each image

as well as how these properties change over time. The user is able to adjust

several parameters important for proper focal adhesion identification. This

system provides a straightforward tool for the global, unbiased assessment of

focal adhesion behavior common in optical microscopy studies. The webserver is

available at: http://faas.bme.unc.edu/.

DOI: 10.12688/f1000research.2-68.v1

PMCID: PMC3752736

PMID: 24358855

1187. Bioinformatics. 2013 Mar 1;29(5):664-5. doi: 10.1093/bioinformatics/btt023. Epub

2013 Jan 17.

SBSI: an extensible distributed software infrastructure for parameter estimation

in systems biology.

Adams R(1), Clark A, Yamaguchi A, Hanlon N, Tsorman N, Ali S, Lebedeva G, Goltsov

A, Sorokin A, Akman OE, Troein C, Millar AJ, Goryanin I, Gilmore S.

Author information:

(1)SynthSys Edinburgh, King's Buildings, University of Edinburgh, Edinburgh, UK.

SUMMARY: Complex computational experiments in Systems Biology, such as fitting

model parameters to experimental data, can be challenging to perform. Not only do

they frequently require a high level of computational power, but the software

needed to run the experiment needs to be usable by scientists with varying levels

of computational expertise, and modellers need to be able to obtain up-to-date

experimental data resources easily. We have developed a software suite, the

Systems Biology Software Infrastructure (SBSI), to facilitate the

parameter-fitting process. SBSI is a modular software suite composed of three

major components: SBSINumerics, a high-performance library containing

parallelized algorithms for performing parameter fitting; SBSIDispatcher, a

middleware application to track experiments and submit jobs to back-end servers;

and SBSIVisual, an extensible client application used to configure optimization

experiments and view results. Furthermore, we have created a plugin

infrastructure to enable project-specific modules to be easily installed. Plugin

developers can take advantage of the existing user-interface and application

framework to customize SBSI for their own uses, facilitated by SBSI's use of

standard data formats.

AVAILABILITY AND IMPLEMENTATION: All SBSI binaries and source-code are freely

available from http://sourceforge.net/projects/sbsi under an Apache 2 open-source

license. The server-side SBSINumerics runs on any Unix-based operating system;

both SBSIVisual and SBSIDispatcher are written in Java and are platform

independent, allowing use on Windows, Linux and Mac OS X. The SBSI project

website at http://www.sbsi.ed.ac.uk provides documentation and tutorials.

DOI: 10.1093/bioinformatics/btt023

PMCID: PMC3582266

PMID: 23329415 [Indexed for MEDLINE]

1188. Fam Cancer. 2013 Mar;12(1):65-73. doi: 10.1007/s10689-012-9577-8.

Feasibility evaluation of an online tool to guide decisions for BRCA1/2 mutation

carriers.

Schackmann EA(1), Munoz DF, Mills MA, Plevritis SK, Kurian AW.

Author information:

(1)Departments of Medicine and of Health Research and Policy, Stanford University

School of Medicine, Stanford, CA 94305-5405, USA.

Women with BRCA1 or BRCA2 (BRCA1/2) mutations face difficult decisions about

managing their high risks of breast and ovarian cancer. We developed an online

tool to guide decisions about cancer risk reduction (available at:

http://brcatool.stanford.edu ), and recruited patients and clinicians to test its

feasibility. We developed questionnaires for women with BRCA1/2 mutations and

clinicians involved in their care, incorporating the System Usability Scale (SUS)

and the Center for Healthcare Evaluation Provider Satisfaction Questionnaire

(CHCE-PSQ). We enrolled BRCA1/2 mutation carriers who were seen by local

physicians or participating in a national advocacy organization, and we enrolled

clinicians practicing at Stanford University and in the surrounding community.

Forty BRCA1/2 mutation carriers and 16 clinicians participated. Both groups found

the tool easy to use, with SUS scores of 82.5-85 on a scale of 1-100; we did not

observe differences according to patient age or gene mutation. General

satisfaction was high, with a mean score of 4.28 (standard deviation (SD) 0.96)

for patients, and 4.38 (SD 0.89) for clinicians, on a scale of 1-5. Most patients

(77.5 %) were comfortable using the tool at home. Both patients and clinicians

agreed that the decision tool could improve patient-doctor encounters (mean

scores 4.50 and 4.69, on a 1-5 scale). Patients and health care providers rated

the decision tool highly on measures of usability and clinical relevance. These

results will guide a larger study of the tool's impact on clinical decisions.

DOI: 10.1007/s10689-012-9577-8

PMCID: PMC4827615

PMID: 23086584 [Indexed for MEDLINE]

1189. Funct Integr Genomics. 2013 Mar;13(1):11-7. doi: 10.1007/s10142-013-0317-4. Epub

2013 Mar 10.

Wheat Zapper: a flexible online tool for colinearity studies in grass genomes.

Alnemer LM(1), Seetan RI, Bassi FM, Chitraranjan C, Helsene A, Loree P, Goshn SB,

Gu YQ, Luo MC, Iqbal MJ, Lazo GR, Denton AM, Kianian SF.

Author information:

(1)Computer Information Systems Department, The University of Jordan, Amman,

11942, Jordan.

In the course of evolution, the genomes of grasses have maintained an observable

degree of gene order conservation. The information available for already

sequenced genomes can be used to predict the gene order of nonsequenced species

by means of comparative colinearity studies. The "Wheat Zapper" application

presented here performs on-demand colinearity analysis between wheat, rice,

Sorghum, and Brachypodium in a simple, time efficient, and flexible manner. This

application was specifically designed to provide plant scientists with a set of

tools, comprising not only synteny inference, but also automated primer design,

intron/exon boundaries prediction, visual representation using the graphic tool

Circos 0.53, and the possibility of downloading FASTA sequences for downstream

applications. Quality of the "Wheat Zapper" prediction was confirmed against the

genome of maize, with good correlation (r > 0.83) observed between the gene order

predicted on the basis of synteny and their actual position on the genome.

Further, the accuracy of "Wheat Zapper" was calculated at 0.65 considering the

"Genome Zipper" application as the "gold" standard. The differences between these

two tools are amply discussed, making the point that "Wheat Zapper" is an

accurate and reliable on-demand tool that is sure to benefit the cereal

scientific community. The Wheat Zapper is available at

http://wge.ndsu.nodak.edu/wheatzapper/ .

DOI: 10.1007/s10142-013-0317-4

PMID: 23474942 [Indexed for MEDLINE]

1190. IEEE Comput Graph Appl. 2013 Mar-Apr;33(2):86-97.

PhotoCloud: Interactive remote exploration of joint 2D and 3D datasets.

Brivio P, Benedetti L, Tarini M, Ponchio F, Cignoni P, Scopigno R.

PhotoCloud is a real-time client-server system for interactive visualization and

exploration of large datasets comprising thousands of calibrated 2D photographs

of a scene and a complex 3D description of the scene. The system isn't tailored

to any specific data acquisition process; it aims at generality and flexibility.

PhotoCloud achieves scalability through a multiresolution dynamic hierarchical

representation of the data, which is remotely stored and accessed by the client

through an efficient cache system. The system includes a compact image browser

and a multiresolution model renderer. PhotoCloud employs iconic visualization of

the images in the 3D space and projects images onto the 3D scene on the fly.

Users can navigate the 2D and 3D spaces with smooth, integrated, seamless

transitions between them. A study with differently skilled users confirms

PhotoCloud's effectiveness and communication power. The Web extras at

http://www.youtube.com/playlist?list=PLHJB2bhmgB7cmYD0ST9CEDMRv1JlX4xPH are

videos demonstrating PhotoCloud, a real-time client-server system for interactive

exploration of large datasets comprising 2D photos and 3D models.

PMID: 24921091 [Indexed for MEDLINE]

1191. IEEE/ACM Trans Comput Biol Bioinform. 2013 Mar-Apr;10(2):436-46. doi:

10.1109/TCBB.2013.21.

Multilabel learning via random label selection for protein subcellular

multilocations prediction.

Wang X(1), Li GZ.

Author information:

(1)Key Laboratory of Embedded System and Service Computing, Ministry of

Education, Department of Control Science and Engineering, Tongji University,

Shanghai 201804, China.

Prediction of protein subcellular localization is an important but challenging

problem, particularly when proteins may simultaneously exist at, or move between,

two or more different subcellular location sites. Most of the existing protein

subcellular localization methods are only used to deal with the single-location

proteins. In the past few years, only a few methods have been proposed to tackle

proteins with multiple locations. However, they only adopt a simple strategy,

that is, transforming the multilocation proteins to multiple proteins with single

location, which does not take correlations among different subcellular locations

into account. In this paper, a novel method named random label selection (RALS)

(multilabel learning via RALS), which extends the simple binary relevance (BR)

method, is proposed to learn from multilocation proteins in an effective and

efficient way. RALS does not explicitly find the correlations among labels, but

rather implicitly attempts to learn the label correlations from data by

augmenting original feature space with randomly selected labels as its additional

input features. Through the fivefold cross-validation test on a benchmark data

set, we demonstrate our proposed method with consideration of label correlations

obviously outperforms the baseline BR method without consideration of label

correlations, indicating correlations among different subcellular locations

really exist and contribute to improvement of prediction performance.

Experimental results on two benchmark data sets also show that our proposed

methods achieve significantly higher performance than some other state-of-the-art

methods in predicting subcellular multilocations of proteins. The prediction web

server is available at >http://levis.tongji.edu.cn:8080/bioinfo/MLPred-Euk/ for

the public usage.

DOI: 10.1109/TCBB.2013.21

PMID: 23929867 [Indexed for MEDLINE]

1192. J Mol Graph Model. 2013 Mar;40:48-53. doi: 10.1016/j.jmgm.2012.12.012. Epub 2013

Jan 5.

ViewMotions Rainbow: a new method to illustrate molecular motions in proteins.

Cockrell GM(1), Kantrowitz ER.

Author information:

(1)Boston College, Department of Chemistry, Merkert Chemistry Center, Chestnut

Hill, MA 02467, USA.

The biological functions of many enzymes are often coupled with significant

conformational changes. The end states of these conformational changes can often

be determined by X-ray crystallography. These X-ray structures are snapshots of

the two extreme conformations in which the macromolecule exists, but the dynamic

movements between the states are not easily visualized in a two-dimensional

illustration. Here we have developed a new method to visualize macromolecular

motions called a ViewMotions Rainbow diagram. These diagrams show the initial and

final states overlaid along with approximately 30 intermediate structures

calculated by linear interpolation of the backbone coordinates of the initial and

final states. This group of structures is then spectrally colored from the

initial structure in blue to the final structure in red. ViewMotions Rainbow

diagrams provide the reader with a much easier way to understand the

macromolecular motions using a single two-dimensional illustration. Since

producing these diagrams requires a number of different software packages, we

have setup the ViewMotions Web Server (http://viewmotions.bc.edu) to

automatically generate these diagrams from two Protein Data Bank files or from

the Database of Macromolecular Movements (http://molmovdb.org).

Copyright © 2013 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.jmgm.2012.12.012

PMCID: PMC3670142

PMID: 23353585 [Indexed for MEDLINE]

1193. Protein Pept Lett. 2013 Mar;20(3):309-17.

Virus-ECC-mPLoc: a multi-label predictor for predicting the subcellular

localization of virus proteins with both single and multiple sites based on a

general form of Chou's pseudo amino acid composition.

Wang X(1), Li GZ, Lu WC.

Author information:

(1)MOE Key Laboratory of Embedded System and Service Computing, Department of

Control Science and Engineering, Tongji University, Shanghai, China.

Protein subcellular localization aims at predicting the location of a protein

within a cell using computational methods. Knowledge of subcellular localization

of viral proteins in a host cell or virus-infected cell is important because it

is closely related to their destructive tendencies and consequences. Prediction

of viral protein subcellular localization is an important but challenging

problem, particularly when proteins may simultaneously exist at, or move between,

two or more different subcellular location sites. Most of the existing protein

subcellular localization methods specialized for viral proteins are only used to

deal with the single-location proteins. To better reflect the characteristics of

multiplex proteins, a new predictor, called Virus-ECC-mPLoc, has been developed

that can be used to deal with the systems containing both singleplex and

multiplex proteins by introducing a powerful multi-label learning approach which

exploits correlations between subcellular locations and by hybridizing the gene

ontology information with the dipeptide composition information. It can be

utilized to identify viral proteins among the following six locations: (1) viral

capsid, (2) host cell membrane, (3) host endoplasmic reticulum, (4) host

cytoplasm, (5) host nucleus, and (6) secreted. Experimental results show that the

overall success rates thus obtained by Virus-ECC-mPLoc are 86.9% for jackknife

test and 87.2% for independent data set test, which are significantly higher than

that by any of the existing predictors. As a user-friendly web-server,

Virus-ECCmPLoc is freely accessible to the public at the web-site

http://levis.tongji.edu.cn:8080/bioinfo/Virus-ECC-mPLoc/.

PMID: 22591474 [Indexed for MEDLINE]

1194. BMC Res Notes. 2013 Feb 23;6:68. doi: 10.1186/1756-0500-6-68.

SHIFT: server for hidden stops analysis in frame-shifted translation.

Gupta A(1), Singh TR.

Author information:

(1)School of Computer Science and IT, DAVV, Indore, M.P., India.

BACKGROUND: Frameshift is one of the three classes of recoding. Frame-shifts lead

to waste of energy, resources and activity of the biosynthetic machinery. In

addition, some peptides synthesized after frame-shifts are probably cytotoxic

which serve as plausible cause for innumerable number of diseases and disorders

such as muscular dystrophies, lysosomal storage disorders, and cancer. Hidden

stop codons occur naturally in coding sequences among all organisms. These codons

are associated with the early termination of translation for incorrect reading

frame selection and help to reduce the metabolic cost related to the frameshift

events. Researchers have identified several consequences of hidden stop codons

and their association with myriad disorders. However the wealth of information

available is speckled and not effortlessly acquiescent to data-mining. To reduce

this gap, this work describes an algorithmic web based tool to study hidden stops

in frameshifted translation for all the lineages through respective genetic code

systems.

FINDINGS: This paper describes SHIFT, an algorithmic web application tool that

provides a user-friendly interface for identifying and analyzing hidden stops in

frameshifted translation of genomic sequences for all available genetic code

systems. We have calculated the correlation between codon usage frequencies and

the plausible contribution of codons towards hidden stops in an off-frame

context. Markovian chains of various order have been used to model hidden stops

in frameshifted peptides and their evolutionary association with naturally

occurring hidden stops. In order to obtain reliable and persuasive estimates for

the naturally occurring and predicted hidden stops statistical measures have been

implemented.

CONCLUSIONS: This paper presented SHIFT, an algorithmic tool that allows

user-friendly exploration, analysis, and visualization of hidden stop codons in

frameshifted translations. It is expected that this web based tool would serve as

a useful complement for analyzing hidden stop codons in all available genetic

code systems. SHIFT is freely available for academic and research purpose at

http://www.nuccore.org/shift/.

DOI: 10.1186/1756-0500-6-68

PMCID: PMC3598200

PMID: 23432998 [Indexed for MEDLINE]

1195. Virol J. 2013 Feb 23;10:62. doi: 10.1186/1743-422X-10-62.

Mutation Reporter Tool: an online tool to interrogate loci of interest, with its

utility demonstrated using hepatitis B virus.

Bell TG(1), Kramvis A.

Author information:

(1)Hepatitis B Virus Diversity Research Programme, School of Clinical Medicine,

Faculty of Health Sciences, University of the Witwatersrand, 7 York Road,

Parktown, Johannesburg 2193, South Africa.

BACKGROUND: An online tool, which extracts and summarises nucleotide or amino

acid sequence data at specified loci of interest, was developed and tested using

the basic core promoter/precore (BCP/PC) region of the hepatitis B virus (HBV).

The tool is aimed at researchers without specialist computer skills.

METHODS: The tool consists of a web-based front-end, with a CGI script, which

runs Python code to generate an output web-page. The Python code searches the

input sequence data for a specified anchor motif, after which it generates

summary tables and graphs of residue and motif distributions.

RESULTS: After the user provides an input file in FASTA format containing aligned

sequence data (nucleotides or amino acids) and specifies an anchor motif at a

known coordinate, the tool summarizes the nucleotides or amino acids at the

specified loci, their frequency and analyzes motif patterns of the loci.The tool

can output a graph that displays the frequency of mutations relative to a

reference sequence. The tool was used to analyze the BCP/PC region of HBV

belonging to subgenotypes A1, A2 and subgenotype D and to serotype HBV. The

"Discovery Mode" ignores conserved loci and assists in identifying potential loci

of interest.

CONCLUSIONS: Although HBV was used to demonstrate the utility of the Mutation

Reporter Tool, the tool has wide application as it is genome-agnostic: nucleotide

or amino acid sequence data from any organism can be processed. Rapid

characterisation of many sequences can be achieved easily when the loci of

interest are known. The tool is available online, without charge, at

http://hvdr.bioinf.wits.ac.za/tools.

DOI: 10.1186/1743-422X-10-62

PMCID: PMC3749809

PMID: 23433201 [Indexed for MEDLINE]

1196. Bioinformatics. 2013 Feb 15;29(4):497-8. doi: 10.1093/bioinformatics/bts705. Epub

2012 Dec 14.

RS-WebPredictor: a server for predicting CYP-mediated sites of metabolism on

drug-like molecules.

Zaretzki J(1), Bergeron C, Huang TW, Rydberg P, Swamidass SJ, Breneman CM.

Author information:

(1)Department of Pathology and Immunology, Washington University School of

Medicine, St. Louis, MO 63130, USA.

SUMMARY: Regioselectivity-WebPredictor (RS-WebPredictor) is a server that

predicts isozyme-specific cytochrome P450 (CYP)-mediated sites of metabolism

(SOMs) on drug-like molecules. Predictions may be made for the promiscuous 2C9,

2D6 and 3A4 CYP isozymes, as well as CYPs 1A2, 2A6, 2B6, 2C8, 2C19 and 2E1.

RS-WebPredictor is the first freely accessible server that predicts the

regioselectivity of the last six isozymes. Server execution time is fast, taking

on average 2s to encode a submitted molecule and 1s to apply a given model,

allowing for high-throughput use in lead optimization projects.

AVAILABILITY: RS-WebPredictor is accessible for free use at

http://reccr.chem.rpi.edu/Software/RS-WebPredictor/

DOI: 10.1093/bioinformatics/bts705

PMCID: PMC3570214

PMID: 23242264 [Indexed for MEDLINE]

1197. Database (Oxford). 2013 Feb 8;2013:bat002. doi: 10.1093/database/bat002. Print

2013.

CrossTope: a curate repository of 3D structures of immunogenic peptide: MHC

complexes.

Sinigaglia M(1), Antunes DA, Rigo MM, Chies JA, Vieira GF.

Author information:

(1)NBLI - Núcleo de Bioinformática do Laboratório de Imunogenética, Department of

Genetics, Universidade Federal do Rio Grande do Sul, 9500 Bento Gonçalves Avenue,

Bldg 43323, Rm 225, 91501-970 Porto Alegre, RS, Brazil.

The CrossTope is a highly curate repository of three-dimensional structures of

peptide:major histocompatibility complex (MHC) class I complexes (pMHC-I). The

complexes hosted by this databank were obtained in protein databases and by

large-scale in silico construction of pMHC-I structures, using a new approach

developed by our group. At this moment, the database contains 182 'non-redundant'

pMHC-I complexes from two human and two murine alleles. A web server provides

interface for database query. The user can download (i) structure coordinate

files and (ii) topological and charges distribution maps images from the T-cell

receptor-interacting surface of pMHC-I complexes. The retrieved structures and

maps can be used to cluster similar epitopes in cross-reactivity approaches, to

analyse viral escape mutations in a structural level or even to improve the

immunogenicity of tumour antigens. Database URL: http://www.crosstope.com.br.

DOI: 10.1093/database/bat002

PMCID: PMC3567486

PMID: 23396301 [Indexed for MEDLINE]

1198. J Mol Biol. 2013 Feb 8;425(3):647-61. doi: 10.1016/j.jmb.2012.11.041. Epub 2012

Dec 7.

Impact of mutations on the allosteric conformational equilibrium.

Weinkam P(1), Chen YC, Pons J, Sali A.

Author information:

(1)Department of Bioengineering and Therapeutic Sciences, University of

California, San Francisco, San Francisco, CA 94158, USA. pweinkam@salilab.org

Allostery in a protein involves effector binding at an allosteric site that

changes the structure and/or dynamics at a distant, functional site. In addition

to the chemical equilibrium of ligand binding, allostery involves a

conformational equilibrium between one protein substate that binds the effector

and a second substate that less strongly binds the effector. We run molecular

dynamics simulations using simple, smooth energy landscapes to sample specific

ligand-induced conformational transitions, as defined by the effector-bound and

effector-unbound protein structures. These simulations can be performed using our

web server (http://salilab.org/allosmod/). We then develop a set of features to

analyze the simulations and capture the relevant thermodynamic properties of the

allosteric conformational equilibrium. These features are based on molecular

mechanics energy functions, stereochemical effects, and structural/dynamic

coupling between sites. Using a machine-learning algorithm on a data set of 10

proteins and 179 mutations, we predict both the magnitude and the sign of the

allosteric conformational equilibrium shift by the mutation; the impact of a

large identifiable fraction of the mutations can be predicted with an average

unsigned error of 1k(B)T. With similar accuracy, we predict the mutation effects

for an 11th protein that was omitted from the initial training and testing of the

machine-learning algorithm. We also assess which calculated thermodynamic

properties contribute most to the accuracy of the prediction.

Copyright © 2012 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jmb.2012.11.041

PMCID: PMC3557769

PMID: 23228330 [Indexed for MEDLINE]

1199. BMC Bioinformatics. 2013 Feb 7;14:44. doi: 10.1186/1471-2105-14-44.

Prediction of vitamin interacting residues in a vitamin binding protein using

evolutionary information.

Panwar B(1), Gupta S, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology (CSIR), Chandigarh,

India.

BACKGROUND: The vitamins are important cofactors in various enzymatic-reactions.

In past, many inhibitors have been designed against vitamin binding pockets in

order to inhibit vitamin-protein interactions. Thus, it is important to identify

vitamin interacting residues in a protein. It is possible to detect

vitamin-binding pockets on a protein, if its tertiary structure is known.

Unfortunately tertiary structures of limited proteins are available. Therefore,

it is important to develop in-silico models for predicting vitamin interacting

residues in protein from its primary structure.

RESULTS: In this study, first we compared protein-interacting residues of

vitamins with other ligands using Two Sample Logo (TSL). It was observed that

ATP, GTP, NAD, FAD and mannose preferred {G,R,K,S,H}, {G,K,T,S,D,N}, {T,G,Y},

{G,Y,W} and {Y,D,W,N,E} residues respectively, whereas vitamins preferred

{Y,F,S,W,T,G,H} residues for the interaction with proteins. Furthermore,

compositional information of preferred and non-preferred residues along with

patterns-specificity was also observed within different vitamin-classes. Vitamins

A, B and B6 preferred {F,I,W,Y,L,V}, {S,Y,G,T,H,W,N,E} and {S,T,G,H,Y,N}

interacting residues respectively. It suggested that protein-binding patterns of

vitamins are different from other ligands, and motivated us to develop separate

predictor for vitamins and their sub-classes. The four different prediction

modules, (i) vitamin interacting residues (VIRs), (ii) vitamin-A interacting

residues (VAIRs), (iii) vitamin-B interacting residues (VBIRs) and (iv)

pyridoxal-5-phosphate (vitamin B6) interacting residues (PLPIRs) have been

developed. We applied various classifiers of SVM, BayesNet, NaiveBayes,

ComplementNaiveBayes, NaiveBayesMultinomial, RandomForest and IBk etc., as

machine learning techniques, using binary and Position-Specific Scoring Matrix

(PSSM) features of protein sequences. Finally, we selected best performing SVM

modules and obtained highest MCC of 0.53, 0.48, 0.61, 0.81 for VIRs, VAIRs,

VBIRs, PLPIRs respectively, using PSSM-based evolutionary information. All the

modules developed in this study have been trained and tested on non-redundant

datasets and evaluated using five-fold cross-validation technique. The

performances were also evaluated on the balanced and different independent

datasets.

CONCLUSIONS: This study demonstrates that it is possible to predict VIRs, VAIRs,

VBIRs and PLPIRs from evolutionary information of protein sequence. In order to

provide service to the scientific community, we have developed web-server and

standalone software VitaPred (http://crdd.osdd.net/raghava/vitapred/).

DOI: 10.1186/1471-2105-14-44

PMCID: PMC3577447

PMID: 23387468 [Indexed for MEDLINE]

1200. J Theor Biol. 2013 Feb 7;318:1-12. doi: 10.1016/j.jtbi.2012.10.033. Epub 2012 Nov

5.

Predicting membrane protein types by incorporating protein topology, domains,

signal peptides, and physicochemical properties into the general form of Chou's

pseudo amino acid composition.

Chen YK(1), Li KB.

Author information:

(1)Institute of Biomedical Informatics, National Yang-Ming University, No.155,

Sec 2, Lih-Nong Street, Taipei, 112, Taiwan, ROC.

The type information of un-annotated membrane proteins provides an important hint

for their biological functions. The experimental determination of membrane

protein types, despite being more accurate and reliable, is not always feasible

due to the costly laboratory procedures, thereby creating a need for the

development of bioinformatics methods. This article describes a novel

computational classifier for the prediction of membrane protein types using

proteins' sequences. The classifier, comprising a collection of one-versus-one

support vector machines, makes use of the following sequence attributes: (1) the

cationic patch sizes, the orientation, and the topology of transmembrane

segments; (2) the amino acid physicochemical properties; (3) the presence of

signal peptides or anchors; and (4) the specific protein motifs. A new voting

scheme was implemented to cope with the multi-class prediction. Both the training

and the testing sequences were collected from SwissProt. Homologous proteins were

removed such that there is no pair of sequences left in the datasets with a

sequence identity higher than 40%. The performance of the classifier was

evaluated by a Jackknife cross-validation and an independent testing experiments.

Results show that the proposed classifier outperforms earlier predictors in

prediction accuracy in seven of the eight membrane protein types. The overall

accuracy was increased from 78.3% to 88.2%. Unlike earlier approaches which

largely depend on position-specific substitution matrices and amino acid

compositions, most of the sequence attributes implemented in the proposed

classifier have supported literature evidences. The classifier has been deployed

as a web server and can be accessed at http://bsaltools.ym.edu.tw/predmpt.

Copyright © 2012 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2012.10.033

PMID: 23137835 [Indexed for MEDLINE]

1201. Zootaxa. 2013 Feb 4;3609:593-600. doi: 10.11646/zootaxa.3609.6.5.

Publishing large DNA sequence data in reduced spaces and lasting formats, in

paper or PDF.

Aguiar AP(1).

Author information:

(1)Universidade Federal do Espírito Santo, Departamento de Ciências Biológicas,

Av. Fernando Ferrari 514, Goiabeiras, Vitoria, ES, 29075-010, Brazil; Email:

aguiar.2@osu.edu.

Scientific publications carry a practical moral duty: they must last. Along that

line of thinking, some methods are proposed to allow economically and

structurally viable publication of DNA sequence data of any size in printed

matter and PDFs. The proposal is primarily aimed at contributing for preserving

information for the future, while allowing authors to avoid information splitting

and complement storage ex situ, that is, in server machines, outside the

publication proper. The technique may also help to solve the impasse between the

ICZN Code requirement that a new nomen be associated to diagnostic characters for

the taxon vs. the phylogenetic definition of taxa, based on cladograms only:

sequence data are characters, and can now be easily and comfortably included in

taxonomic publications, with direct textual mention to their diagnostic sections.

The compression level achieved allows the inclusion of all wanted DNA or RNA

sequences in the same printed matter or PDF publications where the sequences are

cited and discussed. Reduced font sizes, invisible fonts, and original 2D black &

white and color barcodes are illustrated and briefly discussed. The level of data

compression achieved can allow each full page of sequence data, or about 5000

characters, to be precisely coded into a color barcode as small as a square of

1.5 mm. A practical example is provided with Taeniogonalos woodorum Smith

(Hymenoptera, Trigonalidae). Free software to generate publishable barcodes from

txt or FASTA files is provided at www.systaxon.ufes.br/dna.

PMID: 24699621 [Indexed for MEDLINE]

1202. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2013 Feb 1;69(Pt 2):195-200.

doi: 10.1107/S1744309112044387. Epub 2013 Jan 19.

Visualizing ligand molecules in Twilight electron density.

Weichenberger CX(1), Pozharski E, Rupp B.

Author information:

(1)Center for Biomedicine, European Academy of Bozen/Bolzano, Viale Druso 1,

I-39100 Bozen/Bolzano, Italy. christian.weichenberger@eurac.edu

Three-dimensional models of protein structures determined by X-ray

crystallography are based on the interpretation of experimentally derived

electron-density maps. The real-space correlation coefficient (RSCC) provides an

easily comprehensible, objective measure of the residue-based fit of atom

coordinates to electron density. Among protein structure models, protein-ligand

complexes are of special interest, given their contribution to understanding the

molecular underpinnings of biological activity and to drug design. For consumers

of such models, it is not trivial to determine the degree to which

ligand-structure modelling is biased by subjective electron-density

interpretation. A standalone script, Twilight, is presented for the analysis,

visualization and annotation of a pre-filtered set of 2815 protein-ligand

complexes deposited with the PDB as of 15 January 2012 with ligand RSCC values

that are below a threshold of 0.6. It also provides simplified access to the

visualization of any protein-ligand complex available from the PDB and annotated

by the Uppsala Electron Density Server. The script runs on various platforms and

is available for download at http://www.ruppweb.org/twilight/.

DOI: 10.1107/S1744309112044387

PMCID: PMC3564628

PMID: 23385767 [Indexed for MEDLINE]

1203. Bioinformatics. 2013 Feb 1;29(3):384-6. doi: 10.1093/bioinformatics/bts695. Epub

2012 Dec 6.

ChromoZoom: a flexible, fluid, web-based genome browser.

Pak TR(1), Roth FP.

Author information:

(1)Donnelly Centre, University of Toronto, Toronto, ON M5S3E1, Canada.

Current web-based genome browsers require repetitious user input to scroll over

long distances, alter the drawing density of elements or zoom through multiple

orders of magnitude. Generally, either the server or the client is responsible

for the majority of data processing, resulting in either servers having to

receive and handle data relevant only to one user, or clients redundantly

processing widely viewed data. ChromoZoom pre-renders and caches general-use

tracks into tiled images on the server and serves them in an interactive web

interface with inertial scrolling and precise, fluent zooming via the mouse wheel

or trackpad. Custom tracks in several formats can be rendered by client-side code

alongside the pre-rendered tracks, minimizing server load because of

user-specific rendering and eliminating the need to transmit private data.

ChromoZoom thereby enables rapid and simultaneous exploration of curated,

experimental and personal genomic datasets.AVAILABILITY: Human and yeast genome

researchers may browse recent assemblies within ChromoZoom at

http://chromozoom.org/. Source code is available at

http://github.com/rothlab/chromozoom/.

DOI: 10.1093/bioinformatics/bts695

PMCID: PMC3562068

PMID: 23220575 [Indexed for MEDLINE]

1204. Int J Immunogenet. 2013 Feb;40(1):60-5. doi: 10.1111/iji.12030. Epub 2012 Dec 1.

16(th) IHIW: extending the number of resources and bioinformatics analysis for

the investigation of HLA rare alleles.

Gonzalez-Galarza FF(1), Mack SJ, Hollenbach J, Fernandez-Vina M, Setterholm M,

Kempenich J, Marsh SG, Jones AR, Middleton D; HLA Rare Allele Consortium.

Collaborators: Aubrey M, Bengtsson M, Bicalho Mda G, Bohme I, Brown J, Canossi A,

Carcassi C, Carter V, Cate S, Chen-Chung C, Claas F, Collins N, Crowley J, Darke

C, Diaz-Burlinson N, Dormoy A, Dubois V, Dunn P, Fae I, Ficai G, Fischer G, Fleet

A, Fleischhauer K, Frison S, Garbarino L, Godorezky C, Goodman R, Grubic Z,

Guerini FR, Ivanova M, Jindra P, Jobson S, Kaur G, Kesireddy S, Yang EK, Lai S,

Lie B, Ligeiro D, Little AM, Loiseau P, Lokki ML, Longhi E, Malagoli A,

Martinetti M, Masson E, Mattar S, Moraes ME, Morales V, Murgia B, Mytilineos J,

Nesci S, Ozzela G, Papasteriades C, Perasaari J, Pereira S, Poli F, Poole K,

Porfirio B, Poulton K, Rampim G, Ribas F, Richard L, Roelen D, Shaut C, Smith L,

Sprague M, Tavarozzi F, Tavoularis S, Testi M, Tiercy J, Torres M, Tran H,

Umapathy S, Valentini T, Varney M, Vecchiato C, Venigova P, Vidal S, Vidan-Jeras

B, Walkinshaw A, Ward J, Witt C, Wohlwend G, Wroe E, Zhu C, Zunec R.

Author information:

(1)Institute of Integrative Biology, University of Liverpool, Liverpool, UK.

F.Galarza@liv.ac.uk

Continuing a project presented at the 15th International HLA and Immunogenetics

Workshop (IHIWS) on the rarity of HLA alleles, we sought to expand the number of

data sources and bioinformatics tools available in the Allele Frequencies Net

Database website (AFND, www.allelefrequencies.net). In this 16th IHIWS Rare

Alleles project, HLA alleles described in the latest IMGT/HLA Database (release

3.8.0) were queried against different sources including data from registries

(stem cell) and from 74 different laboratories around the world. We demonstrated

that approximately 40% of the alleles officially named in the IMGT/HLA Database

have been reported only once across all different sources. To facilitate the

large-scale analysis of rare alleles, we have produced an online tool called the

Rare Allele Detector that simplifies the detection of alleles that are considered

to be 'very rare', 'rare' or 'frequent'. Tools and associated data can be

accessed via the www.allelefrequencies.net website.

© 2012 Blackwell Publishing Ltd.

DOI: 10.1111/iji.12030

PMID: 23198982 [Indexed for MEDLINE]

1205. J Bioinform Comput Biol. 2013 Feb;11(1):1340008. doi: 10.1142/S0219720013400088.

Epub 2013 Jan 16.

Motif discovery with data mining in 3D protein structure databases: discovery,

validation and prediction of the U-shape zinc binding ("Huf-Zinc") motif.

Maurer-Stroh S(1), Gao H, Han H, Baeten L, Schymkowitz J, Rousseau F, Zhang L,

Eisenhaber F.

Author information:

(1)Bioinformatics Institute-BII, Agency for Science and Technology (A\*STAR), 30

Biopolis Street #07-01, Matrix, 138671 Singapore. sebastianms@bii.a-star.edu.sg

Data mining in protein databases, derivatives from more fundamental protein 3D

structure and sequence databases, has considerable unearthed potential for the

discovery of sequence motif--structural motif--function relationships as the

finding of the U-shape (Huf-Zinc) motif, originally a small student's project,

exemplifies. The metal ion zinc is critically involved in universal biological

processes, ranging from protein-DNA complexes and transcription regulation to

enzymatic catalysis and metabolic pathways. Proteins have evolved a series of

motifs to specifically recognize and bind zinc ions. Many of these, so called

zinc fingers, are structurally independent globular domains with discontinuous

binding motifs made up of residues mostly far apart in sequence. Through a

systematic approach starting from the BRIX structure fragment database, we

discovered that there exists another predictable subset of zinc-binding motifs

that not only have a conserved continuous sequence pattern but also share a

characteristic local conformation, despite being included in totally different

overall folds. While this does not allow general prediction of all Zn binding

motifs, a HMM-based web server, Huf-Zinc, is available for prediction of these

novel, as well as conventional, zinc finger motifs in protein sequences. The

Huf-Zinc webserver can be freely accessed through this URL

(http://mendel.bii.a-star.edu.sg/METHODS/hufzinc/).

DOI: 10.1142/S0219720013400088

PMID: 23427990 [Indexed for MEDLINE]

1206. Nucleic Acids Res. 2013 Feb 1;41(3):1895-900. doi: 10.1093/nar/gks1204. Epub 2012

Dec 16.

McGenus: a Monte Carlo algorithm to predict RNA secondary structures with

pseudoknots.

Bon M(1), Micheletti C, Orland H.

Author information:

(1)Institut de Physique Théorique, CEA Saclay, CNRS URA 2306, 91191

Gif-sur-Yvette, France.

We present McGenus, an algorithm to predict RNA secondary structures with

pseudoknots. The method is based on a classification of RNA structures according

to their topological genus. McGenus can treat sequences of up to 1000 bases and

performs an advanced stochastic search of their minimum free energy structure

allowing for non-trivial pseudoknot topologies. Specifically, McGenus uses a

Monte Carlo algorithm with replica exchange for minimizing a general scoring

function which includes not only free energy contributions for pair stacking,

loop penalties, etc. but also a phenomenological penalty for the genus of the

pairing graph. The good performance of the stochastic search strategy was

successfully validated against TT2NE which uses the same free energy

parametrization and performs exhaustive or partially exhaustive structure search,

albeit for much shorter sequences (up to 200 bases). Next, the method was applied

to other RNA sets, including an extensive tmRNA database, yielding results that

are competitive with existing algorithms. Finally, it is shown that McGenus

highlights possible limitations in the free energy scoring function. The

algorithm is available as a web server at http://ipht.cea.fr/rna/mcgenus.php.

DOI: 10.1093/nar/gks1204

PMCID: PMC3561945

PMID: 23248008 [Indexed for MEDLINE]

1207. OMICS. 2013 Feb;17(2):106-15. doi: 10.1089/omi.2012.0070. Epub 2013 Jan 5.

SCLAP: an adaptive boosting method for predicting subchloroplast localization of

plant proteins.

Saravanan V(1), Lakshmi PT.

Author information:

(1)Centre for Bioinformatics, School of Life Sciences, Pondicherry University,

Pondicherry, India.

Chloroplasts are organelles found in plant system and other photosynthetic

eukaryotes. Since a large number of essential pathways take place in this

organelle, proteins in the chloroplast are considered vital. Therefore, knowledge

about the subchloroplast localization of the chloroplast proteins will provide

precise information in understanding its interaction within the chloroplast. To

address this, an AdaBoost-based prediction system to predict the subchloroplast

localization of chloroplast proteins (SCLAP) was developed. It integrates three

different sequence-based features for prediction, beside the addition of

similarity-based module for significant improvement in prediction performance.

SCLAP achieved an overall accuracy of 89.3% in jackknife cross-validation test

against the benchmark dataset, which was considered highest among existing tools

and equals the SubIdent, and 85.9% accuracy in new error-free dataset. Evaluation

of SCLAP with the independent dataset, five-fold cross-validation, and their

corresponding receiver operator characteristic curve analysis demonstrated the

SCLAP's efficient performance. SCLAP is the webserver implementation of our

algorithm written in PERL. The server can be used to predict the subchloroplast

localization of chloroplast proteins ( http://sclap.bicpu.edu.in/predict.php ).

DOI: 10.1089/omi.2012.0070

PMID: 23289782 [Indexed for MEDLINE]

1208. Protein Pept Lett. 2013 Feb;20(2):218-30.

metaPIS: a sequence-based meta-server for protein interaction site prediction.

Huang J(1), Deng R, Wang J, Wu H, Xiong Y, Wang X.

Author information:

(1)State Key Laboratory of Biocontrol, Sun Yat-sen University, Xingangxi Road,

Guangzhou, People's Republic of China.

The identification of interfaces in protein complexes is effective for the

elucidation of protein function and helps us to understand their roles in

biological processes. With the exponentially growing amount of protein sequence

data, an exploration of new methods that predict protein interaction sites based

solely on sequence information is becoming increasingly urgent. Because a

combination of different methods could produce better results than a single

method, interaction site prediction can be improved through the utilization of

different methods. This paper describes a new method that predicts interaction

sites based on protein sequences by integrating five different algorithms

employing meta-method, Majority Vote and SVMhmm Regression techniques. The

'metaPIS' web-server was implemented for meta-prediction. An evaluation of the

meta-methods using independent datasets revealed that Majority Vote achieved the

highest average Matthews correlation coefficient (0.181) among all the methods

assessed. SVMhmm Regression achieved a lower score but provided a more stable

result. The metaPIS server allows experimental biologists to speculate regarding

protein function by identifying potential interaction sites based on protein

sequence. As a web server, metaPIS is freely accessible to the public at

http://202.116.74.5:84/metapis.

PMID: 22894160 [Indexed for MEDLINE]

1209. Proteins. 2013 Feb;81(2):229-39. doi: 10.1002/prot.24179. Epub 2012 Oct 16.

Toward optimal fragment generations for ab initio protein structure assembly.

Xu D(1), Zhang Y.

Author information:

(1)Department of Computational Medicine and Bioinformatics, University of

Michigan, Ann Arbor, Michigan 48109, USA.

Fragment assembly using structural motifs excised from other solved proteins has

shown to be an efficient method for ab initio protein-structure prediction.

However, how to construct accurate fragments, how to derive optimal restraints

from fragments, and what the best fragment length is are the basic issues yet to

be systematically examined. In this work, we developed a gapless-threading method

to generate position-specific structure fragments. Distance profiles and torsion

angle pairs are then derived from the fragments by statistical consistency

analysis, which achieved comparable accuracy with the machine-learning-based

methods although the fragments were taken from unrelated proteins. When measured

by both accuracies of the derived distance profiles and torsion angle pairs, we

come to a consistent conclusion that the optimal fragment length for structural

assembly is around 10, and at least 100 fragments at each location are needed to

achieve optimal structure assembly. The distant profiles and torsion angle pairs

as derived by the fragments have been successfully used in QUARK for ab initio

protein structure assembly and are provided by the QUARK online server at

http://zhanglab.ccmb. med.umich.edu/QUARK/.

Copyright © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24179

PMCID: PMC3551984

PMID: 22972754 [Indexed for MEDLINE]

1210. Radiat Oncol. 2013 Jan 30;8:23. doi: 10.1186/1748-717X-8-23.

Quality assurance of radiotherapy in the ongoing EORTC 22042-26042 trial for

atypical and malignant meningioma: results from the dummy runs and prospective

individual case Reviews.

Coskun M(1), Straube W, Hurkmans CW, Melidis C, de Haan PF, Villà S, Collette S,

Weber DC.

Author information:

(1)Department of Radiation Oncology, Ankara Oncology Hospital, Ankara, Turkey.

BACKGROUND: The ongoing EORTC 22042-26042 trial evaluates the efficacy of

high-dose radiotherapy (RT) in atypical/malignant meningioma. The results of the

Dummy Run (DR) and prospective Individual Case Review (ICR) were analyzed in this

Quality Assurance (QA) study.

MATERIAL/METHODS: Institutions were requested to submit a protocol compliant

treatment plan for the DR and ICR, respectively. DR-plans (n=12) and ICR-plans

(n=50) were uploaded to the Image-Guided Therapy QA Center of Advanced Technology

Consortium server (http://atc.wustl.edu/) and were assessed prospectively.

RESULTS: Major deviations were observed in 25% (n=3) of DR-plans while no minor

deviations were observed. Major and minor deviations were observed in 22% (n=11)

and 10% (n=5) of the ICR-plans, respectively. Eighteen% of ICRs could not be

analyzed prospectively, as a result of corrupted or late data submission. CTV to

PTV margins were respected in all cases. Deviations were negatively associated

with the number of submitted cases per institution (p=0.0013), with a cutoff of 5

patients per institutions. No association (p=0.12) was observed between DR and

ICR results, suggesting that DR's results did not predict for an improved QA

process in accrued brain tumor patients.

CONCLUSIONS: A substantial number of protocol deviations were observed in this

prospective QA study. The number of cases accrued per institution was a

significant determinant for protocol deviation. These data suggest that

successful DR is not a guarantee for protocol compliance for accrued patients.

Prospective ICRs should be performed to prevent protocol deviations.

DOI: 10.1186/1748-717X-8-23

PMCID: PMC3564920

PMID: 23363568 [Indexed for MEDLINE]

1211. BMC Bioinformatics. 2013 Jan 28;14:30. doi: 10.1186/1471-2105-14-30.

Seq2Ref: a web server to facilitate functional interpretation.

Li W(1), Cong Q, Kinch LN, Grishin NV.

Author information:

(1)Howard Hughes Medical Institute, University of Texas Southwestern Medical

Center, Dallas, TX 75390-9050, USA.

BACKGROUND: The size of the protein sequence database has been exponentially

increasing due to advances in genome sequencing. However, experimentally

characterized proteins only constitute a small portion of the database, such that

the majority of sequences have been annotated by computational approaches.

Current automatic annotation pipelines inevitably introduce errors, making the

annotations unreliable. Instead of such error-prone automatic annotations,

functional interpretation should rely on annotations of 'reference proteins' that

have been experimentally characterized or manually curated.

RESULTS: The Seq2Ref server uses BLAST to detect proteins homologous to a query

sequence and identifies the reference proteins among them. Seq2Ref then reports

publications with experimental characterizations of the identified reference

proteins that might be relevant to the query. Furthermore, a plurality-based

rating system is developed to evaluate the homologous relationships and rank the

reference proteins by their relevance to the query.

CONCLUSIONS: The reference proteins detected by our server will lend insight into

proteins of unknown function and provide extensive information to develop

in-depth understanding of uncharacterized proteins. Seq2Ref is available at:

http://prodata.swmed.edu/seq2ref.

DOI: 10.1186/1471-2105-14-30

PMCID: PMC3573977

PMID: 23356573 [Indexed for MEDLINE]

1212. Implement Sci. 2013 Jan 24;8:13. doi: 10.1186/1748-5908-8-13.

WhatisKT wiki: a case study of a platform for knowledge translation terms and

definitions--descriptive analysis.

McKibbon KA(1), Lokker C, Keepanasseril A, Colquhoun H, Haynes RB, Wilczynski NL.

Author information:

(1)Health Information Research Unit, Dept of Clinical Epidemiology and

Biostatistics, McMaster University, CRL Building, 1280 Main Street West,

Hamilton, ON L8S 4K1, Canada. mckib@mcmaster.ca

BACKGROUND: More than a hundred terms, often with unclear definitions and varying

emphases, are used by health research and practice communities across the world

who are interested in getting the best possible evidence applied (e.g., knowledge

translation, implementation science, diffusion of innovations, and technology

transfer). This makes finding published evidence difficult and can result in

reduced, misinterpreted, or challenging interactions among professionals. Open

dialogue and interaction among various professionals is needed to achieve

consolidation of vocabulary. We use case report methods to describe how we sought

to build an online tool to present the range of terms and facilitate the dialogue

process across groups and disciplines interested in harnessing research evidence

for healthcare.

METHODS: We used a wiki platform from Wikispaces to present the problem of

terminology and make a case and opportunity for collaboration on usage. Wikis are

web sites where communities of users can collaborate online to build content and

discuss progress. We gathered terms related to getting research into practice,

sought published definitions, and posted these on the wiki (WhatisKT

http://whatiskt.wikispaces.com/). We built the wiki in mid-2008 and promoted it

through various groups and publications. This report describes the content of the

site, our promotion efforts, use of the site, and how the site was used for

collaboration up to the end of 2011.

RESULTS: The WhatisKT wiki site now includes more than 120 pages. Traffic to the

site has increased substantially from an average of 200 monthly visits in 2008 to

1700 in 2011. Visitors from 143 countries viewed the wiki in 2011, compared with

12 countries in 2008. However, most use has been limited to short term accesses

of about 40 seconds per visit, and discussion of consolidation and solidifying

terminology is conspicuously absent.

CONCLUSIONS: Although considerable interest exists in the terms and definitions

related to getting research into practice based on increasing numbers of

accesses, use of the WhatisKT wiki site for anything beyond quick lookups was

minimal. Additional efforts must be directed towards increasing the level of

interaction among the members of the site to encourage collaboration on term use.

DOI: 10.1186/1748-5908-8-13

PMCID: PMC3564745

PMID: 23347357 [Indexed for MEDLINE]

1213. BMC Genomics. 2013 Jan 22;14:47. doi: 10.1186/1471-2164-14-47.

Reassessment of the Listeria monocytogenes pan-genome reveals dynamic integration

hotspots and mobile genetic elements as major components of the accessory genome.

Kuenne C(1), Billion A, Mraheil MA, Strittmatter A, Daniel R, Goesmann A,

Barbuddhe S, Hain T, Chakraborty T.

Author information:

(1)Institute of Medical Microbiology, German Centre for Infection Research,

Justus-Liebig-University, D-35392, Giessen, Germany.

BACKGROUND: Listeria monocytogenes is an important food-borne pathogen and model

organism for host-pathogen interaction, thus representing an invaluable target

considering research on the forces governing the evolution of such microbes. The

diversity of this species has not been exhaustively explored yet, as previous

efforts have focused on analyses of serotypes primarily implicated in human

listeriosis. We conducted complete genome sequencing of 11 strains employing 454

GS FLX technology, thereby achieving full coverage of all serotypes including the

first complete strains of serotypes 1/2b, 3c, 3b, 4c, 4d, and 4e. These were

comparatively analyzed in conjunction with publicly available data and assessed

for pathogenicity in the Galleria mellonella insect model.

RESULTS: The species pan-genome of L. monocytogenes is highly stable but open,

suggesting an ability to adapt to new niches by generating or including new

genetic information. The majority of gene-scale differences represented by the

accessory genome resulted from nine hyper variable hotspots, a similar number of

different prophages, three transposons (Tn916, Tn554, IS3-like), and two

mobilizable islands. Only a subset of strains showed CRISPR/Cas bacteriophage

resistance systems of different subtypes, suggesting a supplementary function in

maintenance of chromosomal stability. Multiple phylogenetic branches of the genus

Listeria imply long common histories of strains of each lineage as revealed by a

SNP-based core genome tree highlighting the impact of small mutations for the

evolution of species L. monocytogenes. Frequent loss or truncation of genes

described to be vital for virulence or pathogenicity was confirmed as a recurring

pattern, especially for strains belonging to lineages III and II. New candidate

genes implicated in virulence function were predicted based on functional domains

and phylogenetic distribution. A comparative analysis of small regulatory RNA

candidates supports observations of a differential distribution of trans-encoded

RNA, hinting at a diverse range of adaptations and regulatory impact.

CONCLUSIONS: This study determined commonly occurring hyper variable hotspots and

mobile elements as primary effectors of quantitative gene-scale evolution of

species L. monocytogenes, while gene decay and SNPs seem to represent major

factors influencing long-term evolution. The discovery of common and disparately

distributed genes considering lineages, serogroups, serotypes and strains of

species L. monocytogenes will assist in diagnostic, phylogenetic and functional

research, supported by the comparative genomic GECO-LisDB analysis server

(http://bioinfo.mikrobio.med.uni-giessen.de/geco2lisdb).

DOI: 10.1186/1471-2164-14-47

PMCID: PMC3556495

PMID: 23339658 [Indexed for MEDLINE]

1214. BMC Bioinformatics. 2013 Jan 17;14:22. doi: 10.1186/1471-2105-14-22.

Kerfuffle: a web tool for multi-species gene colocalization analysis.

Aboukhalil R(1), Fendler B, Atwal GS.

Author information:

(1)Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA.

raboukha@cshl.edu

BACKGROUND: The evolutionary pressures that underlie the large-scale functional

organization of the genome are not well understood in eukaryotes. Recent evidence

suggests that functionally similar genes may colocalize (cluster) in the

eukaryotic genome, suggesting the role of chromatin-level gene regulation in

shaping the physical distribution of coordinated genes. However, few of the

bioinformatic tools currently available allow for a systematic study of gene

colocalization across several, evolutionarily distant species. Furthermore, most

tools require the user to input manually curated lists of gene position

information, DNA sequence or gene homology relations between species. With the

growing number of sequenced genomes, there is a need to provide new comparative

genomics tools that can address the analysis of multi-species gene

colocalization.

RESULTS: Kerfuffle is a web tool designed to help discover, visualize, and

quantify the physical organization of genomes by identifying significant gene

colocalization and conservation across the assembled genomes of available species

(currently up to 47, from humans to worms). Kerfuffle only requires the user to

specify a list of human genes and the names of other species of interest. Without

further input from the user, the software queries the e!Ensembl BioMart server to

obtain positional information and discovers homology relations in all genes and

species specified. Using this information, Kerfuffle performs a multi-species

clustering analysis, presents downloadable lists of clustered genes, performs

Monte Carlo statistical significance calculations, estimates how conserved gene

clusters are across species, plots histograms and interactive graphs, allows

users to save their queries, and generates a downloadable visualization of the

clusters using the Circos software. These analyses may be used to further explore

the functional roles of gene clusters by interrogating the enriched molecular

pathways associated with each cluster.

CONCLUSIONS: Kerfuffle is a new, easy-to-use and publicly available tool to aid

our understanding of functional genomics and comparative genomics. This software

allows for flexibility and quick investigations of a user-defined set of genes,

and the results may be saved online for further analysis. Kerfuffle is freely

available at http://atwallab.org/kerfuffle, is implemented in JavaScript (using

jQuery and jsCharts libraries) and PHP 5.2, runs on an Apache server, and stores

data in flat files and an SQLite database.

DOI: 10.1186/1471-2105-14-22

PMCID: PMC3598493

PMID: 23327649 [Indexed for MEDLINE]

1215. BMC Bioinformatics. 2013 Jan 16;14:16. doi: 10.1186/1471-2105-14-16.

The taxonomic name resolution service: an online tool for automated

standardization of plant names.

Boyle B(1), Hopkins N, Lu Z, Raygoza Garay JA, Mozzherin D, Rees T, Matasci N,

Narro ML, Piel WH, McKay SJ, Lowry S, Freeland C, Peet RK, Enquist BJ.

Author information:

(1)Department of Ecology and Evolutionary Biology, University of Arizona Tucson,

P,O, Box 210088, Tucson, AZ 85721, USA. bboyle@email.arizona.edu

BACKGROUND: The digitization of biodiversity data is leading to the widespread

application of taxon names that are superfluous, ambiguous or incorrect,

resulting in mismatched records and inflated species numbers. The ultimate

consequences of misspelled names and bad taxonomy are erroneous scientific

conclusions and faulty policy decisions. The lack of tools for correcting this

'names problem' has become a fundamental obstacle to integrating disparate data

sources and advancing the progress of biodiversity science.

RESULTS: The TNRS, or Taxonomic Name Resolution Service, is an online application

for automated and user-supervised standardization of plant scientific names. The

TNRS builds upon and extends existing open-source applications for name parsing

and fuzzy matching. Names are standardized against multiple reference taxonomies,

including the Missouri Botanical Garden's Tropicos database. Capable of

processing thousands of names in a single operation, the TNRS parses and corrects

misspelled names and authorities, standardizes variant spellings, and converts

nomenclatural synonyms to accepted names. Family names can be included to

increase match accuracy and resolve many types of homonyms. Partial matching of

higher taxa combined with extraction of annotations, accession numbers and

morphospecies allows the TNRS to standardize taxonomy across a broad range of

active and legacy datasets.

CONCLUSIONS: We show how the TNRS can resolve many forms of taxonomic semantic

heterogeneity, correct spelling errors and eliminate spurious names. As a result,

the TNRS can aid the integration of disparate biological datasets. Although the

TNRS was developed to aid in standardizing plant names, its underlying algorithms

and design can be extended to all organisms and nomenclatural codes. The TNRS is

accessible via a web interface at http://tnrs.iplantcollaborative.org/ and as a

RESTful web service and application programming interface. Source code is

available at https://github.com/iPlantCollaborativeOpenSource/TNRS/.

DOI: 10.1186/1471-2105-14-16

PMCID: PMC3554605

PMID: 23324024 [Indexed for MEDLINE]

1216. BMC Bioinformatics. 2013 Jan 16;14:4. doi: 10.1186/1471-2105-14-4.

SyntTax: a web server linking synteny to prokaryotic taxonomy.

Oberto J(1).

Author information:

(1)Université Paris-Sud 11, CNRS, UMR8621, Institut de Génétique et

Microbiologie, 91405, Orsay, France. jacques.oberto@igmors.u-psud.fr

BACKGROUND: The study of the conservation of gene order or synteny constitutes a

powerful methodology to assess the orthology of genomic regions and to predict

functional relationships between genes. The exponential growth of microbial

genomic databases is expected to improve synteny predictions significantly.

Paradoxically, this genomic data plethora, without information on organisms

relatedness, could impair the performance of synteny analysis programs.

RESULTS: In this work, I present SyntTax, a synteny web service designed to take

full advantage of the large amount or archaeal and bacterial genomes by linking

them through taxonomic relationships. SyntTax incorporates a full hierarchical

taxonomic tree allowing intuitive access to all completely sequenced prokaryotes.

Single or multiple organisms can be chosen on the basis of their lineage by

selecting the corresponding rank nodes in the tree. The synteny methodology is

built upon our previously described Absynte algorithm with several additional

improvements.

CONCLUSIONS: SyntTax aims to produce robust syntenies by providing prompt access

to the taxonomic relationships connecting all completely sequenced microbial

genomes. The reduction in redundancy offered by lineage selection presents the

benefit of increasing accuracy while reducing computation time. This web tool was

used to resolve successfully several conserved complex gene clusters described in

the literature. In addition, particular features of SyntTax permit the

confirmation of the involvement of the four components constituting the E. coli

YgjD multiprotein complex responsible for tRNA modification. By analyzing the

clustering evolution of alternative gene fusions, new proteins potentially

interacting with this complex could be proposed. The web service is available at

http://archaea.u-psud.fr/SyntTax.

DOI: 10.1186/1471-2105-14-4

PMCID: PMC3571937

PMID: 23323735 [Indexed for MEDLINE]

1217. Bioinformatics. 2013 Jan 15;29(2):175-81. doi: 10.1093/bioinformatics/bts682.

Epub 2012 Nov 28.

Defining and predicting structurally conserved regions in protein superfamilies.

Huang IK(1), Pei J, Grishin NV.

Author information:

(1)Department of Mathematics, Rice University, Houston, TX 77005, USA.

91huangi@gmail.com

MOTIVATION: The structures of homologous proteins are generally better conserved

than their sequences. This phenomenon is demonstrated by the prevalence of

structurally conserved regions (SCRs) even in highly divergent protein families.

Defining SCRs requires the comparison of two or more homologous structures and is

affected by their availability and divergence, and our ability to deduce

structurally equivalent positions among them. In the absence of multiple

homologous structures, it is necessary to predict SCRs of a protein using

information from only a set of homologous sequences and (if available) a single

structure. Accurate SCR predictions can benefit homology modelling and sequence

alignment.

RESULTS: Using pairwise DaliLite alignments among a set of homologous structures,

we devised a simple measure of structural conservation, termed structural

conservation index (SCI). SCI was used to distinguish SCRs from non-SCRs. A

database of SCRs was compiled from 386 SCOP superfamilies containing 6489 protein

domains. Artificial neural networks were then trained to predict SCRs with

various features deduced from a single structure and homologous sequences.

Assessment of the predictions via a 5-fold cross-validation method revealed that

predictions based on features derived from a single structure perform similarly

to ones based on homologous sequences, while combining sequence and structural

features was optimal in terms of accuracy (0.755) and Matthews correlation

coefficient (0.476). These results suggest that even without information from

multiple structures, it is still possible to effectively predict SCRs for a

protein. Finally, inspection of the structures with the worst predictions

pinpoints difficulties in SCR definitions.

AVAILABILITY: The SCR database and the prediction server can be found at

http://prodata.swmed.edu/SCR.

CONTACT: 91huangi@gmail.com or grishin@chop.swmed.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

Online.

DOI: 10.1093/bioinformatics/bts682

PMCID: PMC3546793

PMID: 23193223 [Indexed for MEDLINE]

1218. BMC Med Inform Decis Mak. 2013 Jan 7;13:5. doi: 10.1186/1472-6947-13-5.

Ancillary study management systems: a review of needs.

Nelson EK(1), Piehler B, Rauch A, Ramsay S, Holman D, Asare S, Asare A, Igra M.

Author information:

(1)LabKey Software, Seattle, WA, USA. eknelson@labkey.com

BACKGROUND: The valuable clinical data, specimens, and assay results collected

during a primary clinical trial or observational study can enable researchers to

answer additional, pressing questions with relatively small investments in new

measurements. However, management of such follow-on, "ancillary" studies is

complex. It requires coordinating across institutions, sites, repositories, and

approval boards, as well as distributing, integrating, and analyzing diverse data

types. General-purpose software systems that simplify the management of ancillary

studies have not yet been explored in the research literature.

METHODS: We have identified requirements for ancillary study management primarily

as part of our ongoing work with a number of large research consortia. These

organizations include the Center for HIV/AIDS Vaccine Immunology (CHAVI), the

Immune Tolerance Network (ITN), the HIV Vaccine Trials Network (HVTN), the U.S.

Military HIV Research Program (MHRP), and the Network for Pancreatic Organ Donors

with Diabetes (nPOD). We also consulted with researchers at a range of other

disease research organizations regarding their workflows and data management

strategies. Lastly, to enhance breadth, we reviewed process documents for

ancillary study management from other organizations.

RESULTS: By exploring characteristics of ancillary studies, we identify

differentiating requirements and scenarios for ancillary study management systems

(ASMSs). Distinguishing characteristics of ancillary studies may include the

collection of additional measurements (particularly new analyses of existing

specimens); the initiation of studies by investigators unaffiliated with the

original study; cross-protocol data pooling and analysis; pre-existing

participant consent; and pre-existing data context and provenance. For an ASMS to

address these characteristics, it would need to address both operational

requirements (e.g., allocating existing specimens) and data management

requirements (e.g., securely distributing and integrating primary and ancillary

data).

CONCLUSIONS: The scenarios and requirements we describe can help guide the

development of systems that make conducting ancillary studies easier, less

expensive, and less error-prone. Given the relatively consistent characteristics

and challenges of ancillary study management, general-purpose ASMSs are likely to

be useful to a wide range of organizations. Using the requirements identified in

this paper, we are currently developing an open-source, general-purpose ASMS

based on LabKey Server (http://www.labkey.org) in collaboration with CHAVI, the

ITN and nPOD.

DOI: 10.1186/1472-6947-13-5

PMCID: PMC3564696

PMID: 23294514 [Indexed for MEDLINE]

1219. Archiving. 2013;2013:74-79.

Improving Software Sustainability: Lessons Learned from Profiles in Science.

Gallagher ME(1).

Author information:

(1)Lister Hill National Center for Biomedical Communications, U.S. National

Library of Medicine; Bethesda, Maryland, USA.

The Profiles in Science® digital library features digitized surrogates of

historical items selected from the archival collections of the U.S. National

Library of Medicine as well as collaborating institutions. In addition, it

contains a database of descriptive, technical and administrative metadata. It

also contains various software components that allow creation of the metadata,

management of the digital items, and access to the items and metadata through the

Profiles in Science Web site [1]. The choices made building the digital library

were designed to maximize the sustainability and long-term survival of all of the

components of the digital library [2]. For example, selecting standard and open

digital file formats rather than proprietary formats increases the sustainability

of the digital files [3]. Correspondingly, using non-proprietary software may

improve the sustainability of the software--either through in-house expertise or

through the open source community. Limiting our digital library software

exclusively to open source software or to software developed in-house has not

been feasible. For example, we have used proprietary operating systems, scanning

software, a search engine, and office productivity software. We did this when

either lack of essential capabilities or the cost-benefit trade-off favored using

proprietary software. We also did so knowing that in the future we would need to

replace or upgrade some of our proprietary software, analogous to migrating from

an obsolete digital file format to a new format as the technological landscape

changes. Since our digital library's start in 1998, all of its software has been

upgraded or replaced, but the digitized items have not yet required migration to

other formats. Technological changes that compelled us to replace proprietary

software included the cost of product licensing, product support, incompatibility

with other software, prohibited use due to evolving security policies, and

product abandonment. Sometimes these changes happen on short notice, so we

continually monitor our library's software for signs of endangerment. We have

attempted to replace proprietary software with suitable in-house or open source

software. When the replacement involves a standalone piece of software with a

nearly equivalent version, such as replacing a commercial HTTP server with an

open source HTTP server, the replacement is straightforward. Recently we replaced

software that functioned not only as our search engine but also as the backbone

of the architecture of our Web site. In this paper, we describe the lessons

learned and the pros and cons of replacing this software with open source

software.

PMCID: PMC4176707

PMID: 25267934

1220. Bioinformatics. 2013 Jan 1;29(1):114-6. doi: 10.1093/bioinformatics/bts636. Epub

2012 Nov 4.

MetaGeneTack: ab initio detection of frameshifts in metagenomic sequences.

Tang S(1), Antonov I, Borodovsky M.

Author information:

(1)School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA.

SUMMARY: Frameshift (FS) prediction is important for analysis and biological

interpretation of metagenomic sequences. Since a genomic context of a short

metagenomic sequence is rarely known, there is not enough data available to

estimate parameters of species-specific statistical models of protein-coding and

non-coding regions. The challenge of ab initio FS detection is, therefore, two

fold: (i) to find a way to infer necessary model parameters and (ii) to identify

positions of frameshifts (if any). Here we describe a new tool, MetaGeneTack,

which uses a heuristic method to estimate parameters of sequence models used in

the FS detection algorithm. It is shown on multiple test sets that the

MetaGeneTack FS detection performance is comparable or better than the one of

earlier developed program FragGeneScan.

AVAILABILITY AND IMPLEMENTATION: MetaGeneTack is available as a web server at

http://exon.gatech.edu/GeneTack/cgi/metagenetack.cgi. Academic users can download

a standalone version of the program from

http://exon.gatech.edu/license\_download.cgi.

DOI: 10.1093/bioinformatics/bts636

PMCID: PMC3530910

PMID: 23129300 [Indexed for MEDLINE]

1221. Bioinformatics. 2013 Jan 1;29(1):62-8. doi: 10.1093/bioinformatics/bts641. Epub

2012 Oct 25.

MycPermCheck: the Mycobacterium tuberculosis permeability prediction tool for

small molecules.

Merget B(1), Zilian D, Müller T, Sotriffer CA.

Author information:

(1)Institute of Pharmacy and Food Chemistry, University of Würzburg, D-97074

Würzburg, Germany.

MOTIVATION: With >8 million new cases in 2010, particularly documented in

developing countries, tuberculosis (TB) is still a highly present pandemic and

often terminal. This is also due to the emergence of antibiotic-resistant strains

(MDR-TB and XDR-TB) of the primary causative TB agent Mycobacterium tuberculosis

(MTB). Efforts to develop new effective drugs against MTB are restrained by the

unique and largely impermeable composition of the mycobacterial cell wall.

RESULTS: Based on a database of antimycobacterial substances (CDD TB), 3815

compounds were classified as active and thus permeable. A data mining approach

was conducted to gather the physico-chemical similarities of these substances and

delimit them from a generic dataset of drug-like molecules. On the basis of the

differences in these datasets, a regression model was generated and implemented

into the online tool MycPermCheck to predict the permeability probability of

small organic compounds.

DISCUSSION: Given the current lack of precise molecular criteria determining

mycobacterial permeability, MycPermCheck represents an unprecedented prediction

tool intended to support antimycobacterial drug discovery. It follows a novel

knowledge-driven approach to estimate the permeability probability of small

organic compounds. As such, MycPermCheck can be used intuitively as an additional

selection criterion for potential new inhibitors against MTB. Based on the

validation results, its performance is expected to be of high practical value for

virtual screening purposes.

AVAILABILITY: The online tool is freely accessible under the URL

http://www.mycpermcheck.aksotriffer.pharmazie.uni-wuerzburg.de

DOI: 10.1093/bioinformatics/bts641

PMID: 23104888 [Indexed for MEDLINE]

1222. Bioinformatics. 2013 Jan 1;29(1):126-8. doi: 10.1093/bioinformatics/bts637. Epub

2012 Oct 25.

Introducing Drugster: a comprehensive and fully integrated drug design, lead and

structure optimization toolkit.

Vlachakis D(1), Tsagrasoulis D, Megalooikonomou V, Kossida S.

Author information:

(1)Bioinformatics and Medical Informatics Team, Biomedical Research Foundation,

Academy of Athens, Athens, Greece.

SUMMARY: Drugster is a fully interactive pipeline designed to break the command

line barrier and introduce a new user-friendly environment to perform drug

design, lead and structure optimization experiments through an efficient

combination of the PDB2PQR, Ligbuilder, Gromacs and Dock suites. Our platform

features a novel workflow that guides the user through each logical step of the

iterative 3D structural optimization setup and drug design process, by providing

a seamless interface to all incorporated packages.

AVAILABILITY: Drugster can be freely downloaded via our dedicated server system

at http://www.bioacademy.gr/bioinformatics/drugster/.

DOI: 10.1093/bioinformatics/bts637

PMID: 23104887 [Indexed for MEDLINE]

1223. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub

2012 Oct 25.

STAR: ultrafast universal RNA-seq aligner.

Dobin A(1), Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P,

Chaisson M, Gingeras TR.

Author information:

(1)Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. dobin@cshl.edu

MOTIVATION: Accurate alignment of high-throughput RNA-seq data is a challenging

and yet unsolved problem because of the non-contiguous transcript structure,

relatively short read lengths and constantly increasing throughput of the

sequencing technologies. Currently available RNA-seq aligners suffer from high

mapping error rates, low mapping speed, read length limitation and mapping

biases.

RESULTS: To align our large (>80 billon reads) ENCODE Transcriptome RNA-seq

dataset, we developed the Spliced Transcripts Alignment to a Reference (STAR)

software based on a previously undescribed RNA-seq alignment algorithm that uses

sequential maximum mappable seed search in uncompressed suffix arrays followed by

seed clustering and stitching procedure. STAR outperforms other aligners by a

factor of >50 in mapping speed, aligning to the human genome 550 million 2 × 76

bp paired-end reads per hour on a modest 12-core server, while at the same time

improving alignment sensitivity and precision. In addition to unbiased de novo

detection of canonical junctions, STAR can discover non-canonical splices and

chimeric (fusion) transcripts, and is also capable of mapping full-length RNA

sequences. Using Roche 454 sequencing of reverse transcription polymerase chain

reaction amplicons, we experimentally validated 1960 novel intergenic splice

junctions with an 80-90% success rate, corroborating the high precision of the

STAR mapping strategy.

AVAILABILITY AND IMPLEMENTATION: STAR is implemented as a standalone C++ code.

STAR is free open source software distributed under GPLv3 license and can be

downloaded from http://code.google.com/p/rna-star/.

DOI: 10.1093/bioinformatics/bts635

PMCID: PMC3530905

PMID: 23104886 [Indexed for MEDLINE]

1224. Bioinformatics. 2013 Jan 1;29(1):8-14. doi: 10.1093/bioinformatics/bts621. Epub

2012 Oct 24.

Simultaneous alignment and clustering of peptide data using a Gibbs sampling

approach.

Andreatta M(1), Lund O, Nielsen M.

Author information:

(1)Center for Biological Sequence Analysis, Technical University of Denmark,

DK-2800 Lyngby, Denmark. massimo@cbs.dtu.dk

MOTIVATION: Proteins recognizing short peptide fragments play a central role in

cellular signaling. As a result of high-throughput technologies, peptide-binding

protein specificities can be studied using large peptide libraries at

dramatically lower cost and time. Interpretation of such large peptide datasets,

however, is a complex task, especially when the data contain multiple receptor

binding motifs, and/or the motifs are found at different locations within

distinct peptides.

RESULTS: The algorithm presented in this article, based on Gibbs sampling,

identifies multiple specificities in peptide data by performing two essential

tasks simultaneously: alignment and clustering of peptide data. We apply the

method to de-convolute binding motifs in a panel of peptide datasets with

different degrees of complexity spanning from the simplest case of pre-aligned

fixed-length peptides to cases of unaligned peptide datasets of variable length.

Example applications described in this article include mixtures of binders to

different MHC class I and class II alleles, distinct classes of ligands for SH3

domains and sub-specificities of the HLA-A\*02:01 molecule.

AVAILABILITY: The Gibbs clustering method is available online as a web server at

http://www.cbs.dtu.dk/services/GibbsCluster.

DOI: 10.1093/bioinformatics/bts621

PMID: 23097419 [Indexed for MEDLINE]

1225. Bioinformation. 2013;9(2):112-5. doi: 10.6026/97320630009112. Epub 2013 Jan 18.

BIRS - Bioterrorism Information Retrieval System.

Tewari AK(1), Rashi, Wadhwa G, Sharma SK, Jain CK.

Author information:

(1)Department of Biotechnology, Jaypee Institute of Information Technology, A-10,

Sector-62, NOIDA, U.P., 201301, India.

Bioterrorism is the intended use of pathogenic strains of microbes to widen

terror in a population. There is a definite need to promote research for

development of vaccines, therapeutics and diagnostic methods as a part of

preparedness to any bioterror attack in the future. BIRS is an open-access

database of collective information on the organisms related to bioterrorism. The

architecture of database utilizes the current open-source technology viz PHP ver

5.3.19, MySQL and IIS server under windows platform for database designing.

Database stores information on literature, generic- information and unique

pathways of about 10 microorganisms involved in bioterrorism. This may serve as a

collective repository to accelerate the drug discovery and vaccines designing

process against such bioterrorist agents (microbes). The available data has been

validated from various online resources and literature mining in order to provide

the user with a comprehensive information system.AVAILABILITY: The database is

freely available at http://www.bioterrorism.biowaves.org.

DOI: 10.6026/97320630009112

PMCID: PMC3563408

PMID: 23390356

1226. BMC Bioinformatics. 2013;14 Suppl 16:S2. doi: 10.1186/1471-2105-14-S16-S2. Epub

2013 Oct 22.

Incorporating substrate sequence motifs and spatial amino acid composition to

identify kinase-specific phosphorylation sites on protein three-dimensional

structures.

Su MG, Lee TY.

BACKGROUND: Protein phosphorylation catalyzed by kinases plays crucial regulatory

roles in cellular processes. Given the high-throughput mass spectrometry-based

experiments, the desire to annotate the catalytic kinases for in vivo

phosphorylation sites has motivated. Thus, a variety of computational methods

have been developed for performing a large-scale prediction of kinase-specific

phosphorylation sites. However, most of the proposed methods solely rely on the

local amino acid sequences surrounding the phosphorylation sites. An increasing

number of three-dimensional structures make it possible to physically investigate

the structural environment of phosphorylation sites.

RESULTS: In this work, all of the experimental phosphorylation sites are mapped

to the protein entries of Protein Data Bank by sequence identity. It resulted in

a total of 4508 phosphorylation sites containing the protein three-dimensional

(3D) structures. To identify phosphorylation sites on protein 3D structures, this

work incorporates support vector machines (SVMs) with the information of linear

motifs and spatial amino acid composition, which is determined for each kinase

group by calculating the relative frequencies of 20 amino acid types within a

specific radial distance from central phosphorylated amino acid residue. After

the cross-validation evaluation, most of the kinase-specific models trained with

the consideration of structural information outperform the models considering

only the sequence information. Furthermore, the independent testing set which is

not included in training set has demonstrated that the proposed method could

provide a comparable performance to other popular tools.

CONCLUSION: The proposed method is shown to be capable of predicting

kinase-specific phosphorylation sites on 3D structures and has been implemented

as a web server which is freely accessible at http://csb.cse.yzu.edu.tw/PhosK3D/.

Due to the difficulty of identifying the kinase-specific phosphorylation sites

with similar sequenced motifs, this work also integrates the 3D structural

information to improve the cross classifying specificity.

DOI: 10.1186/1471-2105-14-S16-S2

PMCID: PMC3853090

PMID: 24564522 [Indexed for MEDLINE]

1227. BMC Bioinformatics. 2013;14 Suppl 16:S10. doi: 10.1186/1471-2105-14-S16-S10. Epub

2013 Oct 22.

ViralPhos: incorporating a recursively statistical method to predict

phosphorylation sites on virus proteins.

Huang KY, Lu CT, Bretaña N, Lee TY, Chang TH.

BACKGROUND: The phosphorylation of virus proteins by host kinases is linked to

viral replication. This leads to an inhibition of normal host-cell functions.

Further elucidation of phosphorylation in virus proteins is required in order to

aid in drug design and treatment. However, only a few studies have investigated

substrate motifs in identifying virus phosphorylation sites. Additionally,

existing bioinformatics tool do not consider potential host kinases that may

initiate the phosphorylation of a virus protein.

RESULTS: 329 experimentally verified phosphorylation fragments on 111 virus

proteins were collected from virPTM. These were clustered into subgroups of

significantly conserved motifs using a recursively statistical method.

Two-layered Support Vector Machines (SVMs) were then applied to train a

predictive model for the identified substrate motifs. The SVM models were

evaluated using a five-fold cross validation which yields an average accuracy of

0.86 for serine, and 0.81 for threonine. Furthermore, the proposed method is

shown to perform at par with three other phosphorylation site prediction tools:

PPSP, KinasePhos 2.0 and GPS 2.1.

CONCLUSION: In this study, we propose a computational method, ViralPhos, which

aims to investigate virus substrate site motifs and identify potential

phosphorylation sites on virus proteins. We identified informative substrate

motifs that matched with several well-studied kinase groups as potential

catalytic kinases for virus protein substrates. The identified substrate motifs

were further exploited to identify potential virus phosphorylation sites. The

proposed method is shown to be capable of predicting virus phosphorylation sites

and has been implemented as a web server http://csb.cse.yzu.edu.tw/ViralPhos/.

DOI: 10.1186/1471-2105-14-S16-S10

PMCID: PMC3853219

PMID: 24564381 [Indexed for MEDLINE]

1228. BMC Bioinformatics. 2013;14 Suppl 18:S2. doi: 10.1186/1471-2105-14-S18-S2. Epub

2013 Nov 5.

Rigidity analysis of protein biological assemblies and periodic crystal

structures.

Jagodzinski F, Clark P, Grant J, Liu T, Monastra S, Streinu I.

BACKGROUND: We initiate in silico rigidity-theoretical studies of biological

assemblies and small crystals for protein structures. The goal is to determine

if, and how, the interactions among neighboring cells and subchains affect the

flexibility of a molecule in its crystallized state. We use experimental X-ray

crystallography data from the Protein Data Bank (PDB). The analysis relies on an

effcient graph-based algorithm. Computational experiments were performed using

new protein rigidity analysis tools available in the new release of our

KINARI-Web server http://kinari.cs.umass.edu.

RESULTS: We provide two types of results: on biological assemblies and on

crystals. We found that when only isolated subchains are considered, structural

and functional information may be missed. Indeed, the rigidity of biological

assemblies is sometimes dependent on the count and placement of hydrogen bonds

and other interactions among the individual subchains of the biological unit.

Similarly, the rigidity of small crystals may be affected by the interactions

between atoms belonging to different unit cells.

CONCLUSION: The rigidity analysis of a single asymmetric unit may not accurately

reflect the protein's behavior in the tightly packed crystal environment. Using

our KINARI software, we demonstrated that additional functional and rigidity

information can be gained by analyzing a protein's biological assembly and/or

crystal structure. However, performing a larger scale study would be

computationally expensive (due to the size of the molecules involved). Overcoming

this limitation will require novel mathematical and computational extensions to

our software.

DOI: 10.1186/1471-2105-14-S18-S2

PMCID: PMC3817814

PMID: 24564201 [Indexed for MEDLINE]

1229. BMC Bioinformatics. 2013;14 Suppl 13:S9. doi: 10.1186/1471-2105-14-S13-S9. Epub

2013 Oct 1.

Dinosolve: a protein disulfide bonding prediction server using context-based

features to enhance prediction accuracy.

Yaseen A, Li Y.

BACKGROUND: Disulfide bonds play an important role in protein folding and

structure stability. Accurately predicting disulfide bonds from protein sequences

is important for modeling the structural and functional characteristics of many

proteins.

METHODS: In this work, we introduce an approach of enhancing disulfide bonding

prediction accuracy by taking advantage of context-based features. We firstly

derive the first-order and second-order mean-force potentials according to the

amino acid environment around the cysteine residues from large number of cysteine

samples. The mean-force potentials are integrated as context-based scores to

estimate the favorability of a cysteine residue in disulfide bonding state as

well as a cysteine pair in disulfide bond connectivity. These context-based

scores are then incorporated as features together with other sequence and

evolutionary information to train neural networks for disulfide bonding state

prediction and connectivity prediction.

RESULTS: The 10-fold cross validated accuracy is 90.8% at residue-level and 85.6%

at protein-level in classifying an individual cysteine residue as bonded or free,

which is around 2% accuracy improvement. The average accuracy for disulfide

bonding connectivity prediction is also improved, which yields overall

sensitivity of 73.42% and specificity of 91.61%.

CONCLUSIONS: Our computational results have shown that the context-based scores

are effective features to enhance the prediction accuracies of both disulfide

bonding state prediction and connectivity prediction. Our disulfide prediction

algorithm is implemented on a web server named "Dinosolve" available at:

http://hpcr.cs.odu.edu/dinosolve.

DOI: 10.1186/1471-2105-14-S13-S9

PMCID: PMC3849605

PMID: 24267383 [Indexed for MEDLINE]

1230. BMC Bioinformatics. 2013;14 Suppl 6:S4. doi: 10.1186/1471-2105-14-S6-S4. Epub

2013 Apr 17.

AllerTOP--a server for in silico prediction of allergens.

Dimitrov I(1), Flower DR, Doytchinova I.

Author information:

(1)Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st,, Sofia,

Bulgaria.

BACKGROUND: Allergy is a form of hypersensitivity to normally innocuous

substances, such as dust, pollen, foods or drugs. Allergens are small antigens

that commonly provoke an IgE antibody response. There are two types of

bioinformatics-based allergen prediction. The first approach follows FAO/WHO

Codex alimentarius guidelines and searches for sequence similarity. The second

approach is based on identifying conserved allergenicity-related linear motifs.

Both approaches assume that allergenicity is a linearly coded property. In the

present study, we applied ACC pre-processing to sets of known allergens,

developing alignment-independent models for allergen recognition based on the

main chemical properties of amino acid sequences.

RESULTS: A set of 684 food, 1,156 inhalant and 555 toxin allergens was collected

from several databases. A set of non-allergens from the same species were

selected to mirror the allergen set. The amino acids in the protein sequences

were described by three z-descriptors (z1, z2 and z3) and by auto- and

cross-covariance (ACC) transformation were converted into uniform vectors. Each

protein was presented as a vector of 45 variables. Five machine learning methods

for classification were applied in the study to derive models for allergen

prediction. The methods were: discriminant analysis by partial least squares

(DA-PLS), logistic regression (LR), decision tree (DT), naïve Bayes (NB) and k

nearest neighbours (kNN). The best performing model was derived by kNN at k = 3.

It was optimized, cross-validated and implemented in a server named AllerTOP,

freely accessible at http://www.pharmfac.net/allertop. AllerTOP also predicts the

most probable route of exposure. In comparison to other servers for allergen

prediction, AllerTOP outperforms them with 94% sensitivity.

CONCLUSIONS: AllerTOP is the first alignment-free server for in silico prediction

of allergens based on the main physicochemical properties of proteins.

Significantly, as well allergenicity AllerTOP is able to predict the route of

allergen exposure: food, inhalant or toxin.

DOI: 10.1186/1471-2105-14-S6-S4

PMCID: PMC3633022

PMID: 23735058 [Indexed for MEDLINE]

1231. BMC Bioinformatics. 2013;14 Suppl 2:S4. doi: 10.1186/1471-2105-14-S2-S4. Epub

2013 Jan 21.

An enhanced computational platform for investigating the roles of regulatory RNA

and for identifying functional RNA motifs.

Chang TH(1), Huang HY, Hsu JB, Weng SL, Horng JT, Huang HD.

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsin-Chu 300, Taiwan.

BACKGROUND: Functional RNA molecules participate in numerous biological

processes, ranging from gene regulation to protein synthesis. Analysis of

functional RNA motifs and elements in RNA sequences can obtain useful information

for deciphering RNA regulatory mechanisms. Our previous work, RegRNA, is widely

used in the identification of regulatory motifs, and this work extends it by

incorporating more comprehensive and updated data sources and analytical

approaches into a new platform.

METHODS AND RESULTS: An integrated web-based system, RegRNA 2.0, has been

developed for comprehensively identifying the functional RNA motifs and sites in

an input RNA sequence. Numerous data sources and analytical approaches are

integrated, and several types of functional RNA motifs and sites can be

identified by RegRNA 2.0: (i) splicing donor/acceptor sites; (ii) splicing

regulatory motifs; (iii) polyadenylation sites; (iv) ribosome binding sites; (v)

rho-independent terminator; (vi) motifs in mRNA 5'-untranslated region (5'UTR)

and 3'UTR; (vii) AU-rich elements; (viii) C-to-U editing sites; (ix)

riboswitches; (x) RNA cis-regulatory elements; (xi) transcriptional regulatory

motifs; (xii) user-defined motifs; (xiii) similar functional RNA sequences; (xiv)

microRNA target sites; (xv) non-coding RNA hybridization sites; (xvi) long stems;

(xvii) open reading frames; (xviii) related information of an RNA sequence. User

can submit an RNA sequence and obtain the predictive results through RegRNA 2.0

web page.

CONCLUSIONS: RegRNA 2.0 is an easy to use web server for identifying regulatory

RNA motifs and functional sites. Through its integrated user-friendly interface,

user is capable of using various analytical approaches and observing results with

graphical visualization conveniently. RegRNA 2.0 is now available at

http://regrna2.mbc.nctu.edu.tw.

DOI: 10.1186/1471-2105-14-S2-S4

PMCID: PMC3549854

PMID: 23369107 [Indexed for MEDLINE]

1232. BMC Genomics. 2013;14 Suppl 3:S3. doi: 10.1186/1471-2164-14-S3-S3. Epub 2013 May

28.

Identifying Mendelian disease genes with the variant effect scoring tool.

Carter H(1), Douville C, Stenson PD, Cooper DN, Karchin R.

Author information:

(1)Department of Biomedical Engineering and Institute for Computational Medicine,

Johns Hopkins University, 3400 N, Charles St, Baltimore, Maryland USA.

BACKGROUND: Whole exome sequencing studies identify hundreds to thousands of rare

protein coding variants of ambiguous significance for human health. Computational

tools are needed to accelerate the identification of specific variants and genes

that contribute to human disease.

RESULTS: We have developed the Variant Effect Scoring Tool (VEST), a supervised

machine learning-based classifier, to prioritize rare missense variants with

likely involvement in human disease. The VEST classifier training set comprised ~

45,000 disease mutations from the latest Human Gene Mutation Database release and

another ~45,000 high frequency (allele frequency >1%) putatively neutral missense

variants from the Exome Sequencing Project. VEST outperforms some of the most

popular methods for prioritizing missense variants in carefully designed holdout

benchmarking experiments (VEST ROC AUC = 0.91, PolyPhen2 ROC AUC = 0.86, SIFT4.0

ROC AUC = 0.84). VEST estimates variant score p-values against a null

distribution of VEST scores for neutral variants not included in the VEST

training set. These p-values can be aggregated at the gene level across multiple

disease exomes to rank genes for probable disease involvement. We tested the

ability of an aggregate VEST gene score to identify candidate Mendelian disease

genes, based on whole-exome sequencing of a small number of disease cases. We

used whole-exome data for two Mendelian disorders for which the causal gene is

known. Considering only genes that contained variants in all cases, the VEST gene

score ranked dihydroorotate dehydrogenase (DHODH) number 2 of 2253 genes in four

cases of Miller syndrome, and myosin-3 (MYH3) number 2 of 2313 genes in three

cases of Freeman Sheldon syndrome.

CONCLUSIONS: Our results demonstrate the potential power gain of aggregating

bioinformatics variant scores into gene-level scores and the general utility of

bioinformatics in assisting the search for disease genes in large-scale exome

sequencing studies. VEST is available as a stand-alone software package at

http://wiki.chasmsoftware.org and is hosted by the CRAVAT web server at

http://www.cravat.us.

DOI: 10.1186/1471-2164-14-S3-S3

PMCID: PMC3665549

PMID: 23819870 [Indexed for MEDLINE]

1233. BMC Genomics. 2013;14 Suppl 3:S2. doi: 10.1186/1471-2164-14-S3-S2. Epub 2013 May

28.

Collective judgment predicts disease-associated single nucleotide variants.

Capriotti E(1), Altman RB, Bromberg Y.

Author information:

(1)Division of Informatics, Department of Pathology, University of Alabama at

Birmingham, Birmingham, AL, USA. emidio@uab.edu

BACKGROUND: In recent years the number of human genetic variants deposited into

the publicly available databases has been increasing exponentially. The latest

version of dbSNP, for example, contains ~50 million validated Single Nucleotide

Variants (SNVs). SNVs make up most of human variation and are often the primary

causes of disease. The non-synonymous SNVs (nsSNVs) result in single amino acid

substitutions and may affect protein function, often causing disease. Although

several methods for the detection of nsSNV effects have already been developed,

the consistent increase in annotated data is offering the opportunity to improve

prediction accuracy.

RESULTS: Here we present a new approach for the detection of disease-associated

nsSNVs (Meta-SNP) that integrates four existing methods: PANTHER, PhD-SNP, SIFT

and SNAP. We first tested the accuracy of each method using a dataset of 35,766

disease-annotated mutations from 8,667 proteins extracted from the SwissVar

database. The four methods reached overall accuracies of 64%-76% with a Matthew's

correlation coefficient (MCC) of 0.38-0.53. We then used the outputs of these

methods to develop a machine learning based approach that discriminates between

disease-associated and polymorphic variants (Meta-SNP). In testing, the combined

method reached 79% overall accuracy and 0.59 MCC, ~3% higher accuracy and ~0.05

higher correlation with respect to the best-performing method. Moreover, for the

hardest-to-define subset of nsSNVs, i.e. variants for which half of the

predictors disagreed with the other half, Meta-SNP attained 8% higher accuracy

than the best predictor.

CONCLUSIONS: Here we find that the Meta-SNP algorithm achieves better performance

than the best single predictor. This result suggests that the methods used for

the prediction of variant-disease associations are orthogonal, encoding different

biologically relevant relationships. Careful combination of predictions from

various resources is therefore a good strategy for the selection of high

reliability predictions. Indeed, for the subset of nsSNVs where all predictors

were in agreement (46% of all nsSNVs in the set), our method reached 87% overall

accuracy and 0.73 MCC. Meta-SNP server is freely accessible at

http://snps.biofold.org/meta-snp.

DOI: 10.1186/1471-2164-14-S3-S2

PMCID: PMC3839641

PMID: 23819846 [Indexed for MEDLINE]

1234. BMC Genomics. 2013;14 Suppl 3:S6. doi: 10.1186/1471-2164-14-S3-S6. Epub 2013 May

28.

WS-SNPs&GO: a web server for predicting the deleterious effect of human protein

variants using functional annotation.

Capriotti E(1), Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R.

Author information:

(1)Division of Informatics, Department of Pathology, University of Alabama at

Birmingham, Birmingham AL, USA. emidio@uab.edu

BACKGROUND: SNPs&GO is a method for the prediction of deleterious Single Amino

acid Polymorphisms (SAPs) using protein functional annotation. In this work, we

present the web server implementation of SNPs&GO (WS-SNPs&GO). The server is

based on Support Vector Machines (SVM) and for a given protein, its input

comprises: the sequence and/or its three-dimensional structure (when available),

a set of target variations and its functional Gene Ontology (GO) terms. The

output of the server provides, for each protein variation, the probabilities to

be associated to human diseases.

RESULTS: The server consists of two main components, including updated versions

of the sequence-based SNPs&GO (recently scored as one of the best algorithms for

predicting deleterious SAPs) and of the structure-based SNPs&GO(3d) programs.

Sequence and structure based algorithms are extensively tested on a large set of

annotated variations extracted from the SwissVar database. Selecting a balanced

dataset with more than 38,000 SAPs, the sequence-based approach achieves 81%

overall accuracy, 0.61 correlation coefficient and an Area Under the Curve (AUC)

of the Receiver Operating Characteristic (ROC) curve of 0.88. For the subset of

~6,600 variations mapped on protein structures available at the Protein Data Bank

(PDB), the structure-based method scores with 84% overall accuracy, 0.68

correlation coefficient, and 0.91 AUC. When tested on a new blind set of

variations, the results of the server are 79% and 83% overall accuracy for the

sequence-based and structure-based inputs, respectively.

CONCLUSIONS: WS-SNPs&GO is a valuable tool that includes in a unique framework

information derived from protein sequence, structure, evolutionary profile, and

protein function. WS-SNPs&GO is freely available at

http://snps.biofold.org/snps-and-go.

DOI: 10.1186/1471-2164-14-S3-S6

PMCID: PMC3665478

PMID: 23819482 [Indexed for MEDLINE]

1235. BMC Syst Biol. 2013;7 Suppl 2:S7. doi: 10.1186/1752-0509-7-S2-S7. Epub 2013 Oct

14.

cGRNB: a web server for building combinatorial gene regulatory networks through

integrated engineering of seed-matching sequence information and gene expression

datasets.

Xu H, Yu H, Tu K, Shi Q, Wei C, Li YY, Li YX.

BACKGROUND: We are witnessing rapid progress in the development of methodologies

for building the combinatorial gene regulatory networks involving both TFs

(Transcription Factors) and miRNAs (microRNAs). There are a few tools available

to do these jobs but most of them are not easy to use and not accessible online.

A web server is especially needed in order to allow users to upload experimental

expression datasets and build combinatorial regulatory networks corresponding to

their particular contexts.

METHODS: In this work, we compiled putative TF-gene, miRNA-gene and TF-miRNA

regulatory relationships from forward-engineering pipelines and curated them as

built-in data libraries. We streamlined the R codes of our two separate

forward-and-reverse engineering algorithms for combinatorial gene regulatory

network construction and formalized them as two major functional modules. As a

result, we released the cGRNB (combinatorial Gene Regulatory Networks Builder): a

web server for constructing combinatorial gene regulatory networks through

integrated engineering of seed-matching sequence information and gene expression

datasets. The cGRNB enables two major network-building modules, one for MPGE

(miRNA-perturbed gene expression) datasets and the other for parallel miRNA/mRNA

expression datasets. A miRNA-centered two-layer combinatorial regulatory cascade

is the output of the first module and a comprehensive genome-wide network

involving all three types of combinatorial regulations (TF-gene, TF-miRNA, and

miRNA-gene) are the output of the second module.

CONCLUSIONS: In this article we propose cGRNB, a web server for building

combinatorial gene regulatory networks through integrated engineering of

seed-matching sequence information and gene expression datasets. Since parallel

miRNA/mRNA expression datasets are rapidly accumulated by the advance of

next-generation sequencing techniques, cGRNB will be very useful tool for

researchers to build combinatorial gene regulatory networks based on expression

datasets. The cGRNB web-server is free and available online at

http://www.scbit.org/cgrnb.

DOI: 10.1186/1752-0509-7-S2-S7

PMCID: PMC3851836

PMID: 24565134 [Indexed for MEDLINE]

1236. Commun Comput Phys. 2013 Jan;13(1):269-284.

DelPhi Web Server: A comprehensive online suite for electrostatic calculations of

biological macromolecules and their complexes.

Sarkar S(1), Witham S(2), Zhang J(1), Zhenirovskyy M(2), Rocchia W(3), Alexov

E(2).

Author information:

(1)Computational Biophysics and Bioinformatics, Department of Physics, Clemson

University, Clemson, SC 29634 ; Department of Computer Science, Clemson

University, Clemson, SC 29634. (2)Computational Biophysics and Bioinformatics,

Department of Physics, Clemson University, Clemson, SC 29634. (3)Italian

Institute of Technology, Genoa, Italy.

Here we report a web server, the DelPhi web server, which utilizes DelPhi program

to calculate electrostatic energies and the corresponding electrostatic potential

and ionic distributions, and dielectric map. The server provides extra services

to fix structural defects, as missing atoms in the structural file and allows for

generation of missing hydrogen atoms. The hydrogen placement and the

corresponding DelPhi calculations can be done with user selected force field

parameters being either Charmm22, Amber98 or OPLS. Upon completion of the

calculations, the user is given option to download fixed and protonated

structural file, together with the parameter and Delphi output files for further

analysis. Utilizing Jmol viewer, the user can see the corresponding structural

file, to manipulate it and to change the presentation. In addition, if the

potential map is requested to be calculated, the potential can be mapped onto the

molecule surface. The DelPhi web server is available from

http://compbio.clemson.edu/delphi\_webserver.

PMCID: PMC3966485

PMID: 24683424

1237. Genome Biol Evol. 2013;5(2):457-67. doi: 10.1093/gbe/evt017.

Improving genome-wide scans of positive selection by using protein isoforms of

similar length.

Villanueva-Cañas JL(1), Laurie S, Albà MM.

Author information:

(1)Evolutionary Genomics Group, Research Programme on Biomedical Informatics

(GRIB), Hospital del Mar Research Institute (IMIM), Universitat Pompeu Fabra

(UPF), Barcelona, Spain.

Large-scale evolutionary studies often require the automated construction of

alignments of a large number of homologous gene families. The majority of

eukaryotic genes can produce different transcripts due to alternative splicing or

transcription initiation, and many such transcripts encode different protein

isoforms. As analyses tend to be gene centered, one single-protein isoform per

gene is selected for the alignment, with the de facto approach being to use the

longest protein isoform per gene (Longest), presumably to avoid including partial

sequences and to maximize sequence information. Here, we show that this approach

is problematic because it increases the number of indels in the alignments due to

the inclusion of nonhomologous regions, such as those derived from

species-specific exons, increasing the number of misaligned positions. With the

aim of ameliorating this problem, we have developed a novel heuristic, Protein

ALignment Optimizer (PALO), which, for each gene family, selects the combination

of protein isoforms that are most similar in length. We examine several

evolutionary parameters inferred from alignments in which the only difference is

the method used to select the protein isoform combination: Longest, PALO, the

combination that results in the highest sequence conservation, and a randomly

selected combination. We observe that Longest tends to overestimate both

nonsynonymous and synonymous substitution rates when compared with PALO, which is

most likely due to an excess of misaligned positions. The estimation of the

fraction of genes that have experienced positive selection by maximum likelihood

is very sensitive to the method of isoform selection employed, both when

alignments are constructed with MAFFT and with Prank(+F). Longest performs better

than a random combination but still estimates up to 3 times more positively

selected genes than the combination showing the highest conservation, indicating

the presence of many false positives. We show that PALO can eliminate the

majority of such false positives and thus that it is a more appropriate approach

for large-scale analyses than Longest. A web server has been set up to facilitate

the use of PALO given a user-defined set of gene families; it is available at

http://evolutionarygenomics.imim.es/palo.

DOI: 10.1093/gbe/evt017

PMCID: PMC3590775

PMID: 23377868 [Indexed for MEDLINE]

1238. Hum Mutat. 2013 Jan;34(1):57-65. doi: 10.1002/humu.22225. Epub 2012 Nov 2.

Predicting the functional, molecular, and phenotypic consequences of amino acid

substitutions using hidden Markov models.

Shihab HA(1), Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN,

Gaunt TR.

Author information:

(1)Bristol Centre for Systems Biomedicine and MRC CAiTE Centre, School of Social

and Community Medicine, University of Bristol, Bristol, United Kingdom.

The rate at which nonsynonymous single nucleotide polymorphisms (nsSNPs) are

being identified in the human genome is increasing dramatically owing to advances

in whole-genome/whole-exome sequencing technologies. Automated methods capable of

accurately and reliably distinguishing between pathogenic and functionally

neutral nsSNPs are therefore assuming ever-increasing importance. Here, we

describe the Functional Analysis Through Hidden Markov Models (FATHMM) software

and server: a species-independent method with optional species-specific

weightings for the prediction of the functional effects of protein missense

variants. Using a model weighted for human mutations, we obtained performance

accuracies that outperformed traditional prediction methods (i.e., SIFT,

PolyPhen, and PANTHER) on two separate benchmarks. Furthermore, in one benchmark,

we achieve performance accuracies that outperform current state-of-the-art

prediction methods (i.e., SNPs&GO and MutPred). We demonstrate that FATHMM can be

efficiently applied to high-throughput/large-scale human and nonhuman genome

sequencing projects with the added benefit of phenotypic outcome associations. To

illustrate this, we evaluated nsSNPs in wheat (Triticum spp.) to identify some of

the important genetic variants responsible for the phenotypic differences

introduced by intense selection during domestication. A Web-based implementation

of FATHMM, including a high-throughput batch facility and a downloadable

standalone package, is available at http://fathmm.biocompute.org.uk.

© 2012 Wiley Periodicals, Inc.

DOI: 10.1002/humu.22225

PMCID: PMC3558800

PMID: 23033316 [Indexed for MEDLINE]

1239. Int J Bioinform Res Appl. 2013;9(3):221-6. doi: 10.1504/IJBRA.2013.053603.

ASFinder: a tool for genome-wide identification of alternatively splicing

transcripts from EST-derived sequences.

Min XJ(1).

Author information:

(1)Center for Applied Chemical Biology, Department of Biological Sciences,

Youngstown State University, Youngstown, OH 44555, USA. xmin@ysu.edu

Expressed Sequence Tags (ESTs) are a rich resource for identifying Alternatively

Splicing (AS) genes. The ASFinder webserver is designed to identify AS isoforms

from EST-derived sequences. Two approaches are implemented in ASFinder. If no

genomic sequences are provided, the server performs a local BLASTN to identify AS

isoforms from ESTs having both ends aligned but an internal segment unaligned.

Otherwise, ASFinder uses SIM4 to map ESTs to the genome, then the overlapping

ESTs that are mapped to the same genomic locus and have internal variable

exon/intron boundaries are identified as AS isoforms. The tool is available at

http://proteomics.ysu.edu/tools/ASFinder.html.

DOI: 10.1504/IJBRA.2013.053603

PMID: 23649736 [Indexed for MEDLINE]

1240. J Comput Aided Mol Des. 2013 Jan;27(1):91-103. doi: 10.1007/s10822-012-9628-0.

Epub 2013 Jan 3.

EuLoc: a web-server for accurately predict protein subcellular localization in

eukaryotes by incorporating various features of sequence segments into the

general form of Chou's PseAAC.

Chang TH(1), Wu LC, Lee TY, Chen SP, Huang HD, Horng JT.

Author information:

(1)Graduate Institute of Biomedical Informatics, Taipei Medical University,

Taipei, Taiwan.

The function of a protein is generally related to its subcellular localization.

Therefore, knowing its subcellular localization is helpful in understanding its

potential functions and roles in biological processes. This work develops a

hybrid method for computationally predicting the subcellular localization of

eukaryotic protein. The method is called EuLoc and incorporates the Hidden Markov

Model (HMM) method, homology search approach and the support vector machines

(SVM) method by fusing several new features into Chou's pseudo-amino acid

composition. The proposed SVM module overcomes the shortcoming of the homology

search approach in predicting the subcellular localization of a protein which

only finds low-homologous or non-homologous sequences in a protein subcellular

localization annotated database. The proposed HMM modules overcome the

shortcoming of SVM in predicting subcellular localizations using few data on

protein sequences. Several features of a protein sequence are considered,

including the sequence-based features, the biological features derived from

PROSITE, NLSdb and Pfam, the post-transcriptional modification features and

others. The overall accuracy and location accuracy of EuLoc are 90.5 and 91.2 %,

respectively, revealing a better predictive performance than obtained elsewhere.

Although the amounts of data of the various subcellular location groups in

benchmark dataset differ markedly, the accuracies of 12 subcellular localizations

of EuLoc range from 82.5 to 100 %, indicating that this tool is much more

balanced than other tools. EuLoc offers a high, balanced predictive power for

each subcellular localization. EuLoc is now available on the web at

http://euloc.mbc.nctu.edu.tw/.

DOI: 10.1007/s10822-012-9628-0

PMID: 23283513 [Indexed for MEDLINE]

1241. Methods Mol Biol. 2013;942:57-68. doi: 10.1007/978-1-62703-119-6\_3.

Designing functional siRNA with reduced off-target effects.

Naito Y(1), Ui-Tei K.

Author information:

(1)Department of Biophysics and Biochemistry, Graduate School of Science,

University of Tokyo, Tokyo, Japan.

RNA interference (RNAi) mediated by small interfering RNA (siRNA) is now widely

used to knock down gene expression in a sequence-specific manner, making it a

powerful tool not only for studying gene functions but also for therapeutic

applications. siRNA decreases the expression level of the intended target gene

with complete complementarity by cleaving its mRNA. However, the efficacy of each

siRNA widely varies depending on its sequence in mammalian cells; only a limited

fraction of randomly designed siRNAs is functional. Moreover, off-target

silencing effects arise when the siRNA has partial complementarity in the seed

region with unintended genes. Here, we describe the rational designing of

functional, off-target effect-reduced siRNAs using siDirect 2.0 Web server

(http://siDirect2.RNAi.jp/). By using the default parameters, siDirect 2.0 can

design at least one qualified siRNA for >94% of human mRNA sequences in the

RefSeq database.

DOI: 10.1007/978-1-62703-119-6\_3

PMID: 23027045 [Indexed for MEDLINE]

1242. Nucleic Acids Res. 2013 Jan;41(Database issue):D824-7. doi: 10.1093/nar/gks1002.

Epub 2012 Nov 30.

2P2Idb: a structural database dedicated to orthosteric modulation of

protein-protein interactions.

Basse MJ(1), Betzi S, Bourgeas R, Bouzidi S, Chetrit B, Hamon V, Morelli X, Roche

P.

Author information:

(1)Laboratory of Integrative Structural and Chemical Biology (iSCB), Centre de

Recherche en Cancérologie de Marseille (CRCM), CNRS UMR 7258, INSERM U 1068,

Institut Paoli-Calmettes, Marseille, France.

Protein-protein interactions are considered as one of the next generation of

therapeutic targets. Specific tools thus need to be developed to tackle this

challenging chemical space. In an effort to derive some common principles from

recent successes, we have built 2P2Idb (freely accessible at

http://2p2idb.cnrs-mrs.fr), a hand-curated structural database dedicated to

protein-protein interactions with known orthosteric modulators. It includes all

interactions for which both the protein-protein and protein-ligand complexes have

been structurally characterized. A web server provides links to related sites of

interest, binding affinity data, pre-calculated structural information about

protein-protein interfaces and 3D interactive views through java applets.

Comparison of interfaces in 2P2Idb to those of representative datasets of

heterodimeric complexes has led to the identification of geometrical parameters

and residue properties to assess the druggability of protein-protein complexes. A

tool is proposed to calculate a series of biophysical and geometrical parameters

that characterize protein-protein interfaces. A large range of descriptors are

computed including, buried accessible surface area, gap volume, non-bonded

contacts, hydrogen-bonds, atom and residue composition, number of segments and

secondary structure contribution. All together the 2P2I database represents a

structural source of information for scientists from academic institutions or

pharmaceutical industries.

DOI: 10.1093/nar/gks1002

PMCID: PMC3531195

PMID: 23203891 [Indexed for MEDLINE]

1243. Nucleic Acids Res. 2013 Jan;41(Database issue):D1021-6. doi: 10.1093/nar/gks1170.

Epub 2012 Nov 27.

GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update.

Schmidt EE(1), Pelz O, Buhlmann S, Kerr G, Horn T, Boutros M.

Author information:

(1)Division Signaling and Functional Genomics, German Cancer Research Center

(DKFZ), D-69120 Heidelberg, Germany.

RNA interference (RNAi) represents a powerful method to systematically study

loss-of-function phenotypes on a large scale with a wide variety of biological

assays, constituting a rich source for the assignment of gene function. The

GenomeRNAi database (http://www.genomernai.org) makes available RNAi phenotype

data extracted from the literature for human and Drosophila. It also provides

RNAi reagent information, along with an assessment as to their efficiency and

specificity. This manuscript describes an update of the database previously

featured in the NAR Database Issue. The new version has undergone a complete

re-design of the user interface, providing an intuitive, flexible framework for

additional functionalities. Screen information and gene-reagent-phenotype

associations are now available for download. The integration with other resources

has been improved by allowing in-links via GenomeRNAi screen IDs, or external

gene or reagent identifiers. A distributed annotation system (DAS) server enables

the visualization of the phenotypes and reagents in the context of a genome

browser. We have added a page listing 'frequent hitters', i.e. genes that show a

phenotype in many screens, which might guide on-going RNAi studies. Structured

annotation guidelines have been established to facilitate consistent curation,

and a submission template for direct submission by data producers is available

for download.

DOI: 10.1093/nar/gks1170

PMCID: PMC3531141

PMID: 23193271 [Indexed for MEDLINE]

1244. Nucleic Acids Res. 2013 Jan;41(Database issue):D700-5. doi: 10.1093/nar/gks1156.

Epub 2012 Nov 27.

Genomicus: five genome browsers for comparative genomics in eukaryota.

Louis A(1), Muffato M, Roest Crollius H.

Author information:

(1)Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, France.

alouis@biologie.ens.fr

Genomicus (http://www.dyogen.ens.fr/genomicus/) is a database and an online tool

that allows easy comparative genomic visualization in >150 eukaryote genomes. It

provides a way to explore spatial information related to gene organization within

and between genomes and temporal relationships related to gene and genome

evolution. For the specific vertebrate phylum, it also provides access to

ancestral gene order reconstructions and conserved non-coding elements

information. We extended the Genomicus database originally dedicated to

vertebrate to four new clades, including plants, non-vertebrate metazoa, protists

and fungi. This visualization tool allows evolutionary phylogenomics analysis and

exploration. Here, we describe the graphical modules of Genomicus and show how it

is capable of revealing differential gene loss and gain, segmental or genome

duplications and study the evolution of a locus through homology relationships.

DOI: 10.1093/nar/gks1156

PMCID: PMC3531091

PMID: 23193262 [Indexed for MEDLINE]

1245. Nucleic Acids Res. 2013 Jan;41(Database issue):D423-9. doi: 10.1093/nar/gks1154.

Epub 2012 Nov 27.

ESTHER, the database of the α/β-hydrolase fold superfamily of proteins: tools to

explore diversity of functions.

Lenfant N(1), Hotelier T, Velluet E, Bourne Y, Marchot P, Chatonnet A.

Author information:

(1)Dynamique Musculaire et Métabolisme, INRA-UM1, Place Viala, 34060 Montpellier,

France.

The ESTHER database, which is freely available via a web server

(http://bioweb.ensam.inra.fr/esther) and is widely used, is dedicated to proteins

with an α/β-hydrolase fold, and it currently contains >30 000 manually curated

proteins. Herein, we report those substantial changes towards improvement that we

have made to improve ESTHER during the past 8 years since our 2004 update. In

particular, we generated 87 new families and increased the coverage of the

UniProt Knowledgebase (UniProtKB). We also renewed the ESTHER website and added

new visualization tools, such as the Overall Table and the Family Tree. We also

address two topics of particular interest to the ESTHER users. First, we explain

how the different enzyme classifications (bacterial lipases, peptidases,

carboxylesterases) used by different communities of users are combined in ESTHER.

Second, we discuss how variations of core architecture or in predicted active

site residues result in a more precise clustering of families, and whether this

strategy provides trustable hints to identify enzyme-like proteins with no

catalytic activity.

DOI: 10.1093/nar/gks1154

PMCID: PMC3531081

PMID: 23193256 [Indexed for MEDLINE]

1246. Nucleic Acids Res. 2013 Jan;41(Database issue):D1159-66. doi:

10.1093/nar/gks1109. Epub 2012 Nov 24.

PIECE: a database for plant gene structure comparison and evolution.

Wang Y(1), You FM, Lazo GR, Luo MC, Thilmony R, Gordon S, Kianian SF, Gu YQ.

Author information:

(1)USDA-Agriculture Research Service, Western Regional Research Center, Albany,

CA 94710, USA.

Gene families often show degrees of differences in terms of exon-intron

structures depending on their distinct evolutionary histories. Comparative

analysis of gene structures is important for understanding their evolutionary and

functional relationships within plant species. Here, we present a comparative

genomics database named PIECE (http://wheat.pw.usda.gov/piece) for Plant Intron

and Exon Comparison and Evolution studies. The database contains all the

annotated genes extracted from 25 sequenced plant genomes. These genes were

classified based on Pfam motifs. Phylogenetic trees were pre-constructed for each

gene category. PIECE provides a user-friendly interface for different types of

searches and a graphical viewer for displaying a gene structure pattern diagram

linked to the resulting bootstrapped dendrogram for each gene family. The gene

structure evolution of orthologous gene groups was determined using the GLOOME,

Exalign and GECA software programs that can be accessed within the database.

PIECE also provides a web server version of the software, GSDraw, for drawing

schematic diagrams of gene structures. PIECE is a powerful tool for comparing

gene sequences and provides valuable insights into the evolution of gene

structure in plant genomes.

DOI: 10.1093/nar/gks1109

PMCID: PMC3531150

PMID: 23180792 [Indexed for MEDLINE]

1247. PLoS Comput Biol. 2013;9(6):e1003088. doi: 10.1371/journal.pcbi.1003088. Epub

2013 Jun 6.

Scrutinizing MHC-I binding peptides and their limits of variation.

Koch CP(1), Perna AM, Pillong M, Todoroff NK, Wrede P, Folkers G, Hiss JA,

Schneider G.

Author information:

(1)ETH Zürich, Department of Chemistry and Applied Biosciences, Institute of

Pharmaceutical Sciences, Zürich, Switzerland.

Designed peptides that bind to major histocompatibility protein I (MHC-I)

allomorphs bear the promise of representing epitopes that stimulate a desired

immune response. A rigorous bioinformatical exploration of sequence patterns

hidden in peptides that bind to the mouse MHC-I allomorph H-2K(b) is presented.

We exemplify and validate these motif findings by systematically dissecting the

epitope SIINFEKL and analyzing the resulting fragments for their binding

potential to H-2K(b) in a thermal denaturation assay. The results demonstrate

that only fragments exclusively retaining the carboxy- or amino-terminus of the

reference peptide exhibit significant binding potential, with the N-terminal

pentapeptide SIINF as shortest ligand. This study demonstrates that sophisticated

machine-learning algorithms excel at extracting fine-grained patterns from

peptide sequence data and predicting MHC-I binding peptides, thereby considerably

extending existing linear prediction models and providing a fresh view on the

computer-based molecular design of future synthetic vaccines. The server for

prediction is available at http://modlab-cadd.ethz.ch (SLiDER tool, MHC-I version

2012).

DOI: 10.1371/journal.pcbi.1003088

PMCID: PMC3674988

PMID: 23754940 [Indexed for MEDLINE]

1248. PLoS Comput Biol. 2013;9(3):e1002977. doi: 10.1371/journal.pcbi.1002977. Epub

2013 Mar 21.

Analysis of physicochemical and structural properties determining HIV-1

coreceptor usage.

Bozek K(1), Lengauer T, Sierra S, Kaiser R, Domingues FS.

Author information:

(1)Max Planck Institute for Computer Science, Saarbrucken, Germany.

bozek@mpi-inf.mpg.de

The relationship of HIV tropism with disease progression and the recent

development of CCR5-blocking drugs underscore the importance of monitoring virus

coreceptor usage. As an alternative to costly phenotypic assays, computational

methods aim at predicting virus tropism based on the sequence and structure of

the V3 loop of the virus gp120 protein. Here we present a numerical descriptor of

the V3 loop encoding its physicochemical and structural properties. The

descriptor allows for structure-based prediction of HIV tropism and

identification of properties of the V3 loop that are crucial for coreceptor

usage. Use of the proposed descriptor for prediction results in a statistically

significant improvement over the prediction based solely on V3 sequence with 3

percentage points improvement in AUC and 7 percentage points in sensitivity at

the specificity of the 11/25 rule (95%). We additionally assessed the predictive

power of the new method on clinically derived 'bulk' sequence data and obtained a

statistically significant improvement in AUC of 3 percentage points over

sequence-based prediction. Furthermore, we demonstrated the capacity of our

method to predict therapy outcome by applying it to 53 samples from patients

undergoing Maraviroc therapy. The analysis of structural features of the loop

informative of tropism indicates the importance of two loop regions and their

physicochemical properties. The regions are located on opposite strands of the

loop stem and the respective features are predominantly charge-, hydrophobicity-

and structure-related. These regions are in close proximity in the bound

conformation of the loop potentially forming a site determinant for the

coreceptor binding. The method is available via server under

http://structure.bioinf.mpi-inf.mpg.de/.

DOI: 10.1371/journal.pcbi.1002977

PMCID: PMC3605109

PMID: 23555214 [Indexed for MEDLINE]

1249. PLoS One. 2013;8(4):e57680. doi: 10.1371/journal.pone.0057680. Epub 2013 Apr 5.

Genome-scale screening of drug-target associations relevant to Ki using a

chemogenomics approach.

Cao DS(1), Liang YZ, Deng Z, Hu QN, He M, Xu QS, Zhou GH, Zhang LX, Deng ZX, Liu

S.

Author information:

(1)Research Center of Modernization of Traditional Chinese Medicines, Central

South University, Changsha, PR China.

The identification of interactions between drugs and target proteins plays a key

role in genomic drug discovery. In the present study, the quantitative binding

affinities of drug-target pairs are differentiated as a measurement to define

whether a drug interacts with a protein or not, and then a chemogenomics

framework using an unbiased set of general integrated features and random forest

(RF) is employed to construct a predictive model which can accurately classify

drug-target pairs. The predictability of the model is further investigated and

validated by several independent validation sets. The built model is used to

predict drug-target associations, some of which were confirmed by comparing

experimental data from public biological resources. A drug-target interaction

network with high confidence drug-target pairs was also reconstructed. This

network provides further insight for the action of drugs and targets. Finally, a

web-based server called PreDPI-Ki was developed to predict drug-target

interactions for drug discovery. In addition to providing a high-confidence list

of drug-target associations for subsequent experimental investigation guidance,

these results also contribute to the understanding of drug-target interactions.

We can also see that quantitative information of drug-target associations could

greatly promote the development of more accurate models. The PreDPI-Ki server is

freely available via: http://sdd.whu.edu.cn/dpiki.

DOI: 10.1371/journal.pone.0057680

PMCID: PMC3618265

PMID: 23577055 [Indexed for MEDLINE]

1250. PLoS One. 2013;8(3):e58759. doi: 10.1371/journal.pone.0058759. Epub 2013 Mar 11.

Comparative GO: a web application for comparative gene ontology and gene

ontology-based gene selection in bacteria.

Fruzangohar M(1), Ebrahimie E, Ogunniyi AD, Mahdi LK, Paton JC, Adelson DL.

Author information:

(1)Centre for Bioinformatics and Computational Genetics, School of Molecular and

Biomedical Science, The University of Adelaide, Adelaide, South Australia,

Australia.

Erratum in

PLoS One. 2015;10(4):e0125537.

The primary means of classifying new functions for genes and proteins relies on

Gene Ontology (GO), which defines genes/proteins using a controlled vocabulary in

terms of their Molecular Function, Biological Process and Cellular Component. The

challenge is to present this information to researchers to compare and discover

patterns in multiple datasets using visually comprehensible and user-friendly

statistical reports. Importantly, while there are many GO resources available for

eukaryotes, there are none suitable for simultaneous, graphical and statistical

comparison between multiple datasets. In addition, none of them supports

comprehensive resources for bacteria. By using Streptococcus pneumoniae as a

model, we identified and collected GO resources including genes, proteins,

taxonomy and GO relationships from NCBI, UniProt and GO organisations. Then, we

designed database tables in PostgreSQL database server and developed a Java

application to extract data from source files and loaded into database

automatically. We developed a PHP web application based on Model-View-Control

architecture, used a specific data structure as well as current and novel

algorithms to estimate GO graphs parameters. We designed different navigation and

visualization methods on the graphs and integrated these into graphical reports.

This tool is particularly significant when comparing GO groups between multiple

samples (including those of pathogenic bacteria) from different sources

simultaneously. Comparing GO protein distribution among up- or down-regulated

genes from different samples can improve understanding of biological pathways,

and mechanism(s) of infection. It can also aid in the discovery of genes

associated with specific function(s) for investigation as a novel vaccine or

therapeutic targets.AVAILABILITY: http://turing.ersa.edu.au/BacteriaGO.

DOI: 10.1371/journal.pone.0058759

PMCID: PMC3594149

PMID: 23536820 [Indexed for MEDLINE]

1251. PLoS One. 2013;8(3):e51307. doi: 10.1371/journal.pone.0051307. Epub 2013 Mar 6.

Protein-protein docking with F(2)Dock 2.0 and GB-rerank.

Chowdhury R(1), Rasheed M, Keidel D, Moussalem M, Olson A, Sanner M, Bajaj C.

Author information:

(1)Department of Computer Science, Institute of Computational Engineering and

Sciences, University of Texas at Austin, Austin, TX, USA.

MOTIVATION: Computational simulation of protein-protein docking can expedite the

process of molecular modeling and drug discovery. This paper reports on our new

F(2) Dock protocol which improves the state of the art in initial stage rigid

body exhaustive docking search, scoring and ranking by introducing improvements

in the shape-complementarity and electrostatics affinity functions, a new

knowledge-based interface propensity term with FFT formulation, a set of novel

knowledge-based filters and finally a solvation energy (GBSA) based reranking

technique. Our algorithms are based on highly efficient data structures including

the dynamic packing grids and octrees which significantly speed up the

computations and also provide guaranteed bounds on approximation error.

RESULTS: The improved affinity functions show superior performance compared to

their traditional counterparts in finding correct docking poses at higher ranks.

We found that the new filters and the GBSA based reranking individually and in

combination significantly improve the accuracy of docking predictions with only

minor increase in computation time. We compared F(2) Dock 2.0 with ZDock 3.0.2

and found improvements over it, specifically among 176 complexes in ZLab

Benchmark 4.0, F(2) Dock 2.0 finds a near-native solution as the top prediction

for 22 complexes; where ZDock 3.0.2 does so for 13 complexes. F(2) Dock 2.0 finds

a near-native solution within the top 1000 predictions for 106 complexes as

opposed to 104 complexes for ZDock 3.0.2. However, there are 17 and 15 complexes

where F(2) Dock 2.0 finds a solution but ZDock 3.0.2 does not and vice versa;

which indicates that the two docking protocols can also complement each other.

AVAILABILITY: The docking protocol has been implemented as a server with a

graphical client (TexMol) which allows the user to manage multiple docking jobs,

and visualize the docked poses and interfaces. Both the server and client are

available for download. Server:

http://www.cs.utexas.edu/~bajaj/cvc/software/f2dock.shtml. Client:

http://www.cs.utexas.edu/~bajaj/cvc/software/f2dockclient.shtml.

DOI: 10.1371/journal.pone.0051307

PMCID: PMC3590208

PMID: 23483883 [Indexed for MEDLINE]

1252. PLoS One. 2013;8(3):e58173. doi: 10.1371/journal.pone.0058173. Epub 2013 Mar 5.

T3\_MM: a Markov model effectively classifies bacterial type III secretion

signals.

Wang Y(1), Sun M, Bao H, White AP.

Author information:

(1)Genomics Research Center, Haerbin Medical University, Harbin, China.

yejun.wang@gmail.com

MOTIVATION: Type III Secretion Systems (T3SSs) play important roles in the

interaction between gram-negative bacteria and their hosts. T3SSs function by

translocating a group of bacterial effector proteins into the host cytoplasm. The

details of specific type III secretion process are yet to be clarified. This

research focused on comparing the amino acid composition within the N-terminal

100 amino acids from type III secretion (T3S) signal sequences or non-T3S

proteins, specifically whether each residue exerts a constraint on residues found

in adjacent positions. We used these comparisons to set up a statistic model to

quantitatively model and effectively distinguish T3S effectors.

RESULTS: In this study, the amino acid composition (Aac) probability profiles

conditional on its sequentially preceding position and corresponding amino acids

were compared between N-terminal sequences of T3S and non-T3S proteins. The

profiles are generally different. A Markov model, namely T3\_MM, was consequently

designed to calculate the total Aac conditional probability difference, i.e., the

likelihood ratio of a sequence being a T3S or a non-T3S protein. With T3\_MM,

known T3S and non-T3S proteins were found to well approximate two distinct normal

distributions. The model could distinguish validated T3S and non-T3S proteins

with a 5-fold cross-validation sensitivity of 83.9% at a specificity of 90.3%.

T3\_MM was also shown to be more robust, accurate, simple, and statistically

quantitative, when compared with other T3S protein prediction models. The high

effectiveness of T3\_MM also indicated the overall Aac difference between

N-termini of T3S and non-T3S proteins, and the constraint of Aac exerted by its

preceding position and corresponding Aac.

AVAILABILITY: An R package for T3\_MM is freely downloadable from:

http://biocomputer.bio.cuhk.edu.hk/softwares/T3\_MM. T3\_MM web server:

http://biocomputer.bio.cuhk.edu.hk/T3DB/T3\_MM.php.

DOI: 10.1371/journal.pone.0058173

PMCID: PMC3589343

PMID: 23472154 [Indexed for MEDLINE]

1253. PLoS One. 2013;8(3):e57731. doi: 10.1371/journal.pone.0057731. Epub 2013 Mar 4.

Alignment of helical membrane protein sequences using AlignMe.

Stamm M(1), Staritzbichler R, Khafizov K, Forrest LR.

Author information:

(1)Computational Structural Biology Group, Max Planck Institute of Biophysics,

Frankfurt am Main, Germany. k46581a@nucc.cc.nagoya-u.ac.jp

Few sequence alignment methods have been designed specifically for integral

membrane proteins, even though these important proteins have distinct

evolutionary and structural properties that might affect their alignments.

Existing approaches typically consider membrane-related information either by

using membrane-specific substitution matrices or by assigning distinct penalties

for gap creation in transmembrane and non-transmembrane regions. Here, we ask

whether favoring matching of predicted transmembrane segments within a standard

dynamic programming algorithm can improve the accuracy of pairwise membrane

protein sequence alignments. We tested various strategies using a specifically

designed program called AlignMe. An updated set of homologous membrane protein

structures, called HOMEP2, was used as a reference for optimizing the gap

penalties. The best of the membrane-protein optimized approaches were then tested

on an independent reference set of membrane protein sequence alignments from the

BAliBASE collection. When secondary structure (S) matching was combined with

evolutionary information (using a position-specific substitution matrix (P)), in

an approach we called AlignMePS, the resultant pairwise alignments were typically

among the most accurate over a broad range of sequence similarities when compared

to available methods. Matching transmembrane predictions (T), in addition to

evolutionary information, and secondary-structure predictions, in an approach

called AlignMePST, generally reduces the accuracy of the alignments of

closely-related proteins in the BAliBASE set relative to AlignMePS, but may be

useful in cases of extremely distantly related proteins for which sequence

information is less informative. The open source AlignMe code is available at

https://sourceforge.net/projects/alignme/, and at http://www.forrestlab.org,

along with an online server and the HOMEP2 data set.

DOI: 10.1371/journal.pone.0057731

PMCID: PMC3587630

PMID: 23469223 [Indexed for MEDLINE]

1254. PLoS One. 2013;8(2):e56742. doi: 10.1371/journal.pone.0056742. Epub 2013 Feb 28.

Towards improved quality of GPCR models by usage of multiple templates and

profile-profile comparison.

Latek D(1), Pasznik P, Carlomagno T, Filipek S.

Author information:

(1)International Institute of Molecular and Cell Biology, Warsaw, Poland.

dlatek@iimcb.gov.pl

G-protein coupled receptors (GPCRs) are targets of nearly one third of the drugs

at the current pharmaceutical market. Despite their importance in many cellular

processes the crystal structures are available for less than 20 unique GPCRs of

the Rhodopsin-like class. Fortunately, even though involved in different

signaling cascades, this large group of membrane proteins has preserved a uniform

structure comprising seven transmembrane helices that allows quite reliable

comparative modeling. Nevertheless, low sequence similarity between the GPCR

family members is still a serious obstacle not only in template selection but

also in providing theoretical models of acceptable quality. An additional level

of difficulty is the prediction of kinks and bulges in transmembrane helices.

Usage of multiple templates and generation of alignments based on sequence

profiles may increase the rate of success in difficult cases of comparative

modeling in which the sequence similarity between GPCRs is exceptionally low.

Here, we present GPCRM, a novel method for fast and accurate generation of GPCR

models using averaging of multiple template structures and profile-profile

comparison. In particular, GPCRM is the first GPCR structure predictor

incorporating two distinct loop modeling techniques: Modeller and Rosetta

together with the filtering of models based on the Z-coordinate. We tested our

approach on all unique GPCR structures determined to date and report its

performance in comparison with other computational methods targeting the

Rhodopsin-like class. We also provide a database of precomputed GPCR models of

the human receptors from that class.AVAILABILITY: GPCRM SERVER AND DATABASE:

http://gpcrm.biomodellab.eu.

DOI: 10.1371/journal.pone.0056742

PMCID: PMC3585245

PMID: 23468878 [Indexed for MEDLINE]

1255. PLoS One. 2013;8(2):e57225. doi: 10.1371/journal.pone.0057225. Epub 2013 Feb 27.

An ensemble method for predicting subnuclear localizations from primary protein

structures.

Han GS(1), Yu ZG, Anh V, Krishnajith AP, Tian YC.

Author information:

(1)School of Mathematics and Computational Science, Xiangtan University, Xiangtan

City, Hunan, China.

BACKGROUND: Predicting protein subnuclear localization is a challenging problem.

Some previous works based on non-sequence information including Gene Ontology

annotations and kernel fusion have respective limitations. The aim of this work

is twofold: one is to propose a novel individual feature extraction method;

another is to develop an ensemble method to improve prediction performance using

comprehensive information represented in the form of high dimensional feature

vector obtained by 11 feature extraction methods.

METHODOLOGY/PRINCIPAL FINDINGS: A novel two-stage multiclass support vector

machine is proposed to predict protein subnuclear localizations. It only

considers those feature extraction methods based on amino acid classifications

and physicochemical properties. In order to speed up our system, an automatic

search method for the kernel parameter is used. The prediction performance of our

method is evaluated on four datasets: Lei dataset, multi-localization dataset,

SNL9 dataset and a new independent dataset. The overall accuracy of prediction

for 6 localizations on Lei dataset is 75.2% and that for 9 localizations on SNL9

dataset is 72.1% in the leave-one-out cross validation, 71.7% for the

multi-localization dataset and 69.8% for the new independent dataset,

respectively. Comparisons with those existing methods show that our method

performs better for both single-localization and multi-localization proteins and

achieves more balanced sensitivities and specificities on large-size and

small-size subcellular localizations. The overall accuracy improvements are 4.0%

and 4.7% for single-localization proteins and 6.5% for multi-localization

proteins. The reliability and stability of our classification model are further

confirmed by permutation analysis.

CONCLUSIONS: It can be concluded that our method is effective and valuable for

predicting protein subnuclear localizations. A web server has been designed to

implement the proposed method. It is freely available at

http://bioinformatics.awowshop.com/snlpred\_page.php.

DOI: 10.1371/journal.pone.0057225

PMCID: PMC3584121

PMID: 23460833 [Indexed for MEDLINE]

1256. PLoS One. 2013;8(2):e56726. doi: 10.1371/journal.pone.0056726. Epub 2013 Feb 15.

Barcode server: a visualization-based genome analysis system.

Mao F(1), Olman V, Wang Y, Xu Y.

Author information:

(1)Computational Systems Biology Lab, Department of Biochemistry and Molecular

Biology and Institute of Bioinformatics, University of Georgia, Athens, Georgia,

United States of America.

We have previously developed a computational method for representing a genome as

a barcode image, which makes various genomic features visually apparent. We have

demonstrated that this visual capability has made some challenging genome

analysis problems relatively easy to solve. We have applied this capability to a

number of challenging problems, including (a) identification of horizontally

transferred genes, (b) identification of genomic islands with special properties

and (c) binning of metagenomic sequences, and achieved highly encouraging

results. These application results inspired us to develop this barcode-based

genome analysis server for public service, which supports the following

capabilities: (a) calculation of the k-mer based barcode image for a provided DNA

sequence; (b) detection of sequence fragments in a given genome with distinct

barcodes from those of the majority of the genome, (c) clustering of provided DNA

sequences into groups having similar barcodes; and (d) homology-based search

using Blast against a genome database for any selected genomic regions deemed to

have interesting barcodes. The barcode server provides a job management

capability, allowing processing of a large number of analysis jobs for

barcode-based comparative genome analyses. The barcode server is accessible at

http://csbl1.bmb.uga.edu/Barcode.

DOI: 10.1371/journal.pone.0056726

PMCID: PMC3574017

PMID: 23457606 [Indexed for MEDLINE]

1257. PLoS One. 2013;8(2):e56499. doi: 10.1371/journal.pone.0056499. Epub 2013 Feb 20.

Hierarchical classification of protein folds using a novel ensemble classifier.

Lin C(1), Zou Y, Qin J, Liu X, Jiang Y, Ke C, Zou Q.

Author information:

(1)School of Information Science and Technology, Xiamen University, Xiamen,

China. chenlin@xmu.edu.cn

The analysis of biological information from protein sequences is important for

the study of cellular functions and interactions, and protein fold recognition

plays a key role in the prediction of protein structures. Unfortunately, the

prediction of protein fold patterns is challenging due to the existence of

compound protein structures. Here, we processed the latest release of the

Structural Classification of Proteins (SCOP, version 1.75) database and exploited

novel techniques to impressively increase the accuracy of protein fold

classification. The techniques proposed in this paper include ensemble

classifying and a hierarchical framework, in the first layer of which similar or

redundant sequences were deleted in two manners; a set of base classifiers, fused

by various selection strategies, divides the input into seven classes; in the

second layer of which, an analogous ensemble method is adopted to predict all

protein folds. To our knowledge, it is the first time all protein folds can be

intelligently detected hierarchically. Compared with prior studies, our

experimental results demonstrated the efficiency and effectiveness of our

proposed method, which achieved a success rate of 74.21%, which is much higher

than results obtained with previous methods (ranging from 45.6% to 70.5%). When

applied to the second layer of classification, the prediction accuracy was in the

range between 23.13% and 46.05%. This value, which may not be remarkably high, is

scientifically admirable and encouraging as compared to the relatively low counts

of proteins from most fold recognition programs. The web server Hierarchical

Protein Fold Prediction (HPFP) is available at

http://datamining.xmu.edu.cn/software/hpfp.

DOI: 10.1371/journal.pone.0056499

PMCID: PMC3577917

PMID: 23437146 [Indexed for MEDLINE]

1258. PLoS One. 2013;8(2):e56833. doi: 10.1371/journal.pone.0056833. Epub 2013 Feb 19.

Predicting the binding patterns of hub proteins: a study using yeast protein

interaction networks.

Andorf CM(1), Honavar V, Sen TZ.

Author information:

(1)Department of Computer Science, Iowa State University, Ames, Iowa, United

States of America.

BACKGROUND: Protein-protein interactions are critical to elucidating the role

played by individual proteins in important biological pathways. Of particular

interest are hub proteins that can interact with large numbers of partners and

often play essential roles in cellular control. Depending on the number of

binding sites, protein hubs can be classified at a structural level as

singlish-interface hubs (SIH) with one or two binding sites, or

multiple-interface hubs (MIH) with three or more binding sites. In terms of

kinetics, hub proteins can be classified as date hubs (i.e., interact with

different partners at different times or locations) or party hubs (i.e.,

simultaneously interact with multiple partners).

METHODOLOGY: Our approach works in 3 phases: Phase I classifies if a protein is

likely to bind with another protein. Phase II determines if a protein-binding

(PB) protein is a hub. Phase III classifies PB proteins as singlish-interface

versus multiple-interface hubs and date versus party hubs. At each stage, we use

sequence-based predictors trained using several standard machine learning

techniques.

CONCLUSIONS: Our method is able to predict whether a protein is a protein-binding

protein with an accuracy of 94% and a correlation coefficient of 0.87; identify

hubs from non-hubs with 100% accuracy for 30% of the data; distinguish date

hubs/party hubs with 69% accuracy and area under ROC curve of 0.68; and SIH/MIH

with 89% accuracy and area under ROC curve of 0.84. Because our method is based

on sequence information alone, it can be used even in settings where reliable

protein-protein interaction data or structures of protein-protein complexes are

unavailable to obtain useful insights into the functional and evolutionary

characteristics of proteins and their interactions.

AVAILABILITY: We provide a web server for our three-phase approach:

http://hybsvm.gdcb.iastate.edu.

DOI: 10.1371/journal.pone.0056833

PMCID: PMC3576370

PMID: 23431393 [Indexed for MEDLINE]

1259. PLoS One. 2013;8(2):e55844. doi: 10.1371/journal.pone.0055844. Epub 2013 Feb 7.

iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating

position specific amino acid propensity into pseudo amino acid composition.

Xu Y(1), Ding J, Wu LY, Chou KC.

Author information:

(1)Department of Information and Computer Science, University of Science and

Technology Beijing, Beijing, China. yxu@gordonlifescience.org

Posttranslational modifications (PTMs) of proteins are responsible for sensing

and transducing signals to regulate various cellular functions and signaling

events. S-nitrosylation (SNO) is one of the most important and universal PTMs.

With the avalanche of protein sequences generated in the post-genomic age, it is

highly desired to develop computational methods for timely identifying the exact

SNO sites in proteins because this kind of information is very useful for both

basic research and drug development. Here, a new predictor, called iSNO-PseAAC,

was developed for identifying the SNO sites in proteins by incorporating the

position-specific amino acid propensity (PSAAP) into the general form of pseudo

amino acid composition (PseAAC). The predictor was implemented using the

conditional random field (CRF) algorithm. As a demonstration, a benchmark dataset

was constructed that contains 731 SNO sites and 810 non-SNO sites. To reduce the

homology bias, none of these sites were derived from the proteins that had

[Formula: see text] pairwise sequence identity to any other. It was observed that

the overall cross-validation success rate achieved by iSNO-PseAAC in identifying

nitrosylated proteins on an independent dataset was over 90%, indicating that the

new predictor is quite promising. Furthermore, a user-friendly web-server for

iSNO-PseAAC was established at http://app.aporc.org/iSNO-PseAAC/, by which users

can easily obtain the desired results without the need to follow the mathematical

equations involved during the process of developing the prediction method. It is

anticipated that iSNO-PseAAC may become a useful high throughput tool for

identifying the SNO sites, or at the very least play a complementary role to the

existing methods in this area.

DOI: 10.1371/journal.pone.0055844

PMCID: PMC3567014

PMID: 23409062 [Indexed for MEDLINE]

1260. PLoS One. 2013;8(1):e54032. doi: 10.1371/journal.pone.0054032. Epub 2013 Jan 17.

TFPP: an SVM-based tool for recognizing flagellar proteins in Trypanosoma brucei.

Zhang X(1), Shen Y, Ding G, Tian Y, Liu Z, Li B, Wang Y, Jiang C.

Author information:

(1)Department of Bioinformatics, the School of Life Sciences and Technology,

Tongji University, Shanghai, China. zhangxb@tongji.edu.cn

Trypanosoma brucei is a unicellular flagellated eukaryotic parasite that causes

African trypanosomiasis in human and domestic animals with devastating health and

economic consequences. Recent studies have revealed the important roles of the

single flagellum of T. brucei in many aspects, especially that the flagellar

motility is required for the viability of the bloodstream form T. brucei,

suggesting that impairment of the flagellar function may provide a promising cure

for African sleeping sickness. Knowing the flagellum proteome is crucial to study

the molecular mechanism of the flagellar functions. Here we present a novel

computational method for identifying flagellar proteins in T. brucei, called

trypanosome flagellar protein predictor (TFPP). TFPP was developed based on a

list of selected discriminating features derived from protein sequences, and

could predict flagellar proteins with ∼92% specificity at a ∼84% sensitivity

rate. Applied to the whole T. brucei proteome, TFPP reveals 811 more flagellar

proteins with high confidence, suggesting that the flagellar proteome covers ∼10%

of the whole proteome. Comparison of the expression profiles of the whole T.

brucei proteome at three typical life cycle stages found that ∼45% of the

flagellar proteins were significantly changed in expression levels between the

three life cycle stages, indicating life cycle stage-specific regulation of

flagellar functions in T. brucei. Overall, our study demonstrated that TFPP is

highly effective in identifying flagellar proteins and could provide

opportunities to study the trypanosome flagellar proteome systematically.

Furthermore, the web server for TFPP can be freely accessed at

http:/wukong.tongji.edu.cn/tfpp.

DOI: 10.1371/journal.pone.0054032

PMCID: PMC3547966

PMID: 23349782 [Indexed for MEDLINE]

1261. PLoS One. 2013;8(1):e53685. doi: 10.1371/journal.pone.0053685. Epub 2013 Jan 9.

mirTarPri: improved prioritization of microRNA targets through incorporation of

functional genomics data.

Wang P(1), Ning S, Wang Q, Li R, Ye J, Zhao Z, Li Y, Huang T, Li X.

Author information:

(1)College of Bioinformatics Science and Technology, Harbin Medical University,

Harbin, China.

MicroRNAs (miRNAs) are a class of small (19-25 nt) non-coding RNAs. This

important class of gene regulator downregulates gene expression through

sequence-specific binding to the 3'untranslated regions (3'UTRs) of target mRNAs.

Several computational target prediction approaches have been developed for

predicting miRNA targets. However, the predicted target lists often have high

false positive rates. To construct a workable target list for subsequent

experimental studies, we need novel approaches to properly rank the candidate

targets from traditional methods. We performed a systematic analysis of

experimentally validated miRNA targets using functional genomics data, and found

significant functional associations between genes that were targeted by the same

miRNA. Based on this finding, we developed a miRNA target prioritization method

named mirTarPri to rank the predicted target lists from commonly used target

prediction methods. Leave-one-out cross validation has proved to be successful in

identifying known targets, achieving an AUC score up to 0. 84. Validation in

high-throughput data proved that mirTarPri was an unbiased method. Applying

mirTarPri to prioritize results of six commonly used target prediction methods

allowed us to find more positive targets at the top of the prioritized candidate

list. In comparison with other methods, mirTarPri had an outstanding performance

in gold standard and CLIP data. mirTarPri was a valuable method to improve the

efficacy of current miRNA target prediction methods. We have also developed a

web-based server for implementing mirTarPri method, which is freely accessible at

http://bioinfo.hrbmu.edu.cn/mirTarPri.

DOI: 10.1371/journal.pone.0053685

PMCID: PMC3541237

PMID: 23326485 [Indexed for MEDLINE]

1262. PLoS One. 2013;8(1):e53235. doi: 10.1371/journal.pone.0053235. Epub 2013 Jan 7.

Prediction and analysis of antibody amyloidogenesis from sequences.

Liaw C(1), Tung CW, Ho SY.

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu, Taiwan.

Antibody amyloidogenesis is the aggregation of soluble proteins into amyloid

fibrils that is one of major causes of the failures of humanized antibodies. The

prediction and prevention of antibody amyloidogenesis are helpful for restoring

and enhancing therapeutic effects. Due to a large number of possible germlines,

the existing method is not practical to predict sequences of novel germlines,

which establishes individual models for each known germline. This study proposes

a first automatic and across-germline prediction method (named AbAmyloid) capable

of predicting antibody amyloidogenesis from sequences. Since the amyloidogenesis

is determined by a whole sequence of an antibody rather than germline-dependent

properties such as mutated residues, this study assess three types of

germline-independent sequence features (amino acid composition, dipeptide

composition and physicochemical properties). AbAmyloid using a Random Forests

classifier with dipeptide composition performs well on a data set of 12

germlines. The within- and across-germline prediction accuracies are 83.10% and

83.33% using Jackknife tests, respectively, and the novel-germline prediction

accuracy using a leave-one-germline-out test is 72.22%. A thorough analysis of

sequence features is conducted to identify informative properties for further

providing insights to antibody amyloidogenesis. Some identified informative

physicochemical properties are amphiphilicity, hydrophobicity, reverse turn,

helical structure, isoelectric point, net charge, mutability, coil, turn, linker,

nuclear protein, etc. Additionally, the numbers of ubiquitylation sites in

amyloidogenic and non-amyloidogenic antibodies are found to be significantly

different. It reveals that antibodies less likely to be ubiquitylated tend to be

amyloidogenic. The method AbAmyloid capable of automatically predicting antibody

amyloidogenesis of novel germlines is implemented as a publicly available web

server at http://iclab.life.nctu.edu.tw/abamyloid.

DOI: 10.1371/journal.pone.0053235

PMCID: PMC3538782

PMID: 23308169 [Indexed for MEDLINE]

1263. Proteins. 2013 Jan;81(1):149-62. doi: 10.1002/prot.24172. Epub 2012 Sep 29.

CAD-score: a new contact area difference-based function for evaluation of protein

structural models.

Olechnovič K(1), Kulberkytė E, Venclovas C.

Author information:

(1)Institute of Biotechnology, Vilnius University, Graičiūno 8, LT-02241 Vilnius,

Lithuania.

Evaluation of protein models against the native structure is essential for the

development and benchmarking of protein structure prediction methods. Although a

number of evaluation scores have been proposed to date, many aspects of model

assessment still lack desired robustness. In this study we present CAD-score, a

new evaluation function quantifying differences between physical contacts in a

model and the reference structure. The new score uses the concept of

residue-residue contact area difference (CAD) introduced by Abagyan and Totrov (J

Mol Biol 1997; 268:678-685). Contact areas, the underlying basis of the score,

are derived using the Voronoi tessellation of protein structure. The newly

introduced CAD-score is a continuous function, confined within fixed limits, free

of any arbitrary thresholds or parameters. The built-in logic for treatment of

missing residues allows consistent ranking of models of any degree of

completeness. We tested CAD-score on a large set of diverse models and compared

it to GDT-TS, a widely accepted measure of model accuracy. Similarly to GDT-TS,

CAD-score showed a robust performance on single-domain proteins, but displayed a

stronger preference for physically more realistic models. Unlike GDT-TS, the new

score revealed a balanced assessment of domain rearrangement, removing the

necessity for different treatment of single-domain, multi-domain, and

multi-subunit structures. Moreover, CAD-score makes it possible to assess the

accuracy of inter-domain or inter-subunit interfaces directly. In addition, the

approach offers an alternative to the superposition-based model clustering. The

CAD-score implementation is available both as a web server and a standalone

software package at http://www.ibt.lt/bioinformatics/cad-score/.

Copyright © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24172

PMID: 22933340 [Indexed for MEDLINE]

1264. Proteins. 2013 Jan;81(1):140-8. doi: 10.1002/prot.24171. Epub 2012 Sep 26.

Swfoldrate: predicting protein folding rates from amino acid sequence with

sliding window method.

Cheng X(1), Xiao X, Wu ZC, Wang P, Lin WZ.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China.

Protein folding is the process by which a protein processes from its denatured

state to its specific biologically active conformation. Understanding the

relationship between sequences and the folding rates of proteins remains an

important challenge. Most previous methods of predicting protein folding rate

require the tertiary structure of a protein as an input. In this study, the

long-range and short-range contact in protein were used to derive extended

version of the pseudo amino acid composition based on sliding window method. This

method is capable of predicting the protein folding rates just from the amino

acid sequence without the aid of any structural class information. We

systematically studied the contributions of individual features to folding rate

prediction. The optimal feature selection procedures are adopted by means of

combining the forward feature selection and sequential backward selection method.

Using the jackknife cross validation test, the method was demonstrated on the

large dataset. The predictor was achieved on the basis of multitudinous

physicochemical features and statistical features from protein using nonlinear

support vector machine (SVM) regression model, the method obtained an excellent

agreement between predicted and experimentally observed folding rates of

proteins. The correlation coefficient is 0.9313 and the standard error is 2.2692.

The prediction server is freely available at

http://www.jci-bioinfo.cn/swfrate/input.jsp.

Copyright © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24171

PMID: 22933332 [Indexed for MEDLINE]

1265. Proteins. 2013 Jan;81(1):119-31. doi: 10.1002/prot.24167. Epub 2012 Sep 26.

3Drefine: consistent protein structure refinement by optimizing hydrogen bonding

network and atomic-level energy minimization.

Bhattacharya D(1), Cheng J.

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, Missouri

65211, USA.

One of the major limitations of computational protein structure prediction is the

deviation of predicted models from their experimentally derived true, native

structures. The limitations often hinder the possibility of applying

computational protein structure prediction methods in biochemical assignment and

drug design that are very sensitive to structural details. Refinement of these

low-resolution predicted models to high-resolution structures close to the native

state, however, has proven to be extremely challenging. Thus, protein structure

refinement remains a largely unsolved problem. Critical assessment of techniques

for protein structure prediction (CASP) specifically indicated that most

predictors participating in the refinement category still did not consistently

improve model quality. Here, we propose a two-step refinement protocol, called

3Drefine, to consistently bring the initial model closer to the native structure.

The first step is based on optimization of hydrogen bonding (HB) network and the

second step applies atomic-level energy minimization on the optimized model using

a composite physics and knowledge-based force fields. The approach has been

evaluated on the CASP benchmark data and it exhibits consistent improvement over

the initial structure in both global and local structural quality measures.

3Drefine method is also computationally inexpensive, consuming only few minutes

of CPU time to refine a protein of typical length (300 residues). 3Drefine web

server is freely available at http://sysbio.rnet.missouri.edu/3Drefine/.

Copyright © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24167

PMCID: PMC3634918

PMID: 22927229 [Indexed for MEDLINE]

1266. Sci Rep. 2013;3:1607. doi: 10.1038/srep01607.

Computational approach for designing tumor homing peptides.

Sharma A(1), Kapoor P, Gautam A, Chaudhary K, Kumar R, Chauhan JS, Tyagi A,

Raghava GP.

Author information:

(1)Bioinformatics Centre, CSIR-Institute of Microbial Technology,

Chandigarh-160036, India.

Tumor homing peptides are small peptides that home specifically to tumor and

tumor associated microenvironment i.e. tumor vasculature, after systemic

delivery. Keeping in mind the huge therapeutic importance of these peptides, we

have made an attempt to analyze and predict tumor homing peptides. It was

observed that certain types of residues are preferred in tumor homing peptides.

Therefore, we developed support vector machine based models for predicting tumor

homing peptides using amino acid composition and binary profiles of peptides.

Amino acid composition, dipeptide composition and binary profile-based models

achieved a maximum accuracy of 86.56%, 82.03%, and 84.19% respectively. These

methods have been implemented in a user-friendly web server, TumorHPD. We

anticipate that this method will be helpful to design novel tumor homing

peptides. TumorHPD web server is freely accessible at

http://crdd.osdd.net/raghava/tumorhpd/.

DOI: 10.1038/srep01607

PMCID: PMC3617442

PMID: 23558316 [Indexed for MEDLINE]

1267. Syst Biol. 2013 Jan 1;62(1):157-61. doi: 10.1093/sysbio/sys069. Epub 2012 Aug 3.

IKey+: a new single-access key generation web service.

Burguiere T(1), Causse F, Ung V, Vignes-Lebbe R.

Author information:

(1)Laboratoire Informatique et Systématique, Museum National d'Histoire

Naturelle, Département Histoire de la Terre, UMR 7207 - C.N.R.S. - M.N.H.N. -

U.P.M.C., Paris, France. thomas.burguiere@upmc.fr

Single-access keys are a major tool for biologists who need to identify

specimens. The construction process of these keys is particularly complex

(especially if the input data set is large) so having an automatic single-access

key generation tool is essential. As part of the European project ViBRANT, our

aim was to develop such a tool as a web service, thus allowing end-users to

integrate it directly into their workflow. IKey+generates single-access keys on

demand, for single users or research institutions. It receives user input data

(using the standard SDD format), accepts several key-generation parameters

(affecting the key topology and representation), and supports several output

formats. IKey+is freely available (sources and binary packages) at

www.identificationkey.fr. Furthermore, it is deployed on our server and can be

queried (for testing purposes) through a simple web client also available at

www.identificationkey.fr (last accessed 13 August 2012). Finally, a client plugin

will be integrated to the Scratchpads biodiversity networking tool

(scratchpads.eu).

DOI: 10.1093/sysbio/sys069

PMID: 22863681 [Indexed for MEDLINE]

1268. Proceedings (IEEE Int Conf Bioinformatics Biomed). 2012 Dec 31;2012:1-7.

An Accurate Scalable Template-based Alignment Algorithm.

Gardner DP(1), Xu W(2), Miranker DP(3), Ozer S(4), Cannone JJ(5), Gutell RR(5).

Author information:

(1)Institute for Cellular and Molecular Biology The University of Texas at Austin

Austin, TX, USA dpgardn@utexas.edu. (2)Texas Advanced Computing Center The

University of Texas at Austin Austin, TX, USA xwj@tacc.utexas.edu. (3)Department

of Computer Sciences The University of Texas at Austin Austin, TX, USA

miranker@cs.utexas.edu. (4)Microsoft Corporation Redmond, WA, USA

stuarto@microsoft.com. (5)Center for Computational Biology and Bioinformatics The

University of Texas at Austin Austin, TX, USA.

The rapid determination of nucleic acid sequences is increasing the number of

sequences that are available. Inherent in a template or seed alignment is the

culmination of structural and functional constraints that are selecting those

mutations that are viable during the evolution of the RNA. While we might not

understand these structural and functional, template-based alignment programs

utilize the patterns of sequence conservation to encapsulate the characteristics

of viable RNA sequences that are aligned properly. We have developed a program

that utilizes the different dimensions of information in rCAD, a large RNA

informatics resource, to establish a profile for each position in an alignment.

The most significant include sequence identity and column composition in

different phylogenetic taxa. We have compared our methods with a maximum of eight

alternative alignment methods on different sets of 16S and 23S rRNA sequences

with sequence percent identities ranging from 50% to 100%. The results showed

that CRWAlign outperformed the other alignment methods in both speed and

accuracy. A web-based alignment server is available at

http://www.rna.ccbb.utexas.edu/SAE/2F/CRWAlign.

DOI: 10.1109/BIBM.2012.6392676

PMCID: PMC3999978

PMID: 24772376

1269. J Chem Inf Model. 2012 Dec 21;52(12):3341-51. doi: 10.1021/ci300328y. Epub 2012

Nov 16.

Improving the selectivity of antimicrobial peptides from anuran skin.

Kamech N(1), Vukičević D, Ladram A, Piesse C, Vasseur J, Bojović V, Simunić J,

Juretić D.

Author information:

(1)Université Pierre et Marie Curie-Paris 06, Equipe Biogenèse des signaux

peptidiques, ER3, 7 Quai Saint-Bernard, 75252 Paris cedex 05, France.

Anuran skin is known to be a rich source of antimicrobial peptides although their

therapeutic potential is often limited due to their toxicity against mammalian

cells. The analysis of structure-activity relationships among anuran

antimicrobial peptides provided the parameters to construct the "Mutator" tool

for improving their selectivity for bacterial cells, by suggesting appropriate

point substitutions. Double substitution analogues [K2, K16] of the Xenopus

tropicalis peptide XT-7 and [I2, K19] of the Ascaphus truei peptide ascaphin-8

were predicted by this tool to have an increased 'therapeutic index' (TI =

HC(50)/MIC for erythrocytes with respect to bacteria) > 80. The mutated peptides

were synthesized and respectively found to have experimental TI values > 130 for

S. aureus or E. coli, a considerable improvement with respect to TI < 37 for the

parent compounds. Circular dichroism studies of the mutated peptides suggested

this may in part be due to variations in the α-helical structure. For P.

aeruginosa, which is more resistant to XT-7, the TI increased in the mutated

peptide from 5 to >270, also due to a significant improvement in minimal

inhibitory concentration. We have shown that the Mutator tool is capable of

suggesting limited variations in natural anuran peptides capable of increasing

peptide selectivity, by decreasing toxicity against mammalian erythrocytes, in

general without compromising antibacterial activity. The tool is freely available

on the Mutator Web server at http://split4.pmfst.hr/mutator/.

DOI: 10.1021/ci300328y

PMID: 23094651 [Indexed for MEDLINE]

1270. J Proteomics. 2012 Dec 21;77:321-8. doi: 10.1016/j.jprot.2012.09.006. Epub 2012

Sep 20.

Identification of mycobacterial membrane proteins and their types using

over-represented tripeptide compositions.

Ding C(1), Yuan LF, Guo SH, Lin H, Chen W.

Author information:

(1)Key Laboratory for NeuroInformation of Ministry of Education, Center of

Bioinformatics, School of Life Science and Technology, University of Electronic

Science and Technology of China, Chengdu 610054, China.

Mycobacterium can cause many serious diseases, such as tuberculosis and leprosy.

Its membrane proteins play a critical role for multidrug-resistance and its

tenacious survival ability. Knowing the types of membrane proteins will provide

novel insights into understanding their functions and facilitate drug target

discovery. In this study, a novel method was developed for predicting

mycobacterial membrane protein and their types by using over-represented

tripeptides. A total of 295 non-membrane proteins and 274 membrane proteins were

collected to evaluate the performance of proposed method. The results of

jackknife cross-validation test show that our method achieves an overall accuracy

of 93.0% in discriminating between mycobacterial membrane proteins and

mycobacterial non-membrane proteins and an overall accuracy of 93.1% in

classifying mycobacterial membrane protein types. By comparing with other

methods, the proposed method showed excellent predictive performance. Based on

the proposed method, we built a predictor, called MycoMemSVM, which is freely

available at http://lin.uestc.edu.cn/server/MycoMemSVM. It is anticipated that

MycoMemSVM will become a useful tool for the annotation of mycobacterial membrane

proteins and the development of anti-mycobacterium drug design.

Copyright © 2012 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.jprot.2012.09.006

PMID: 23000219 [Indexed for MEDLINE]

1271. BMC Genomics. 2012 Dec 20;13:715. doi: 10.1186/1471-2164-13-715.

CpGAVAS, an integrated web server for the annotation, visualization, analysis,

and GenBank submission of completely sequenced chloroplast genome sequences.

Liu C(1), Shi L, Zhu Y, Chen H, Zhang J, Lin X, Guan X.

Author information:

(1)Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences,

Peking Union Medical College, 151 MaLianWa North Road, Haidian District, Beijing

100191, People's Republic of China. cliu@implad.ac.cn

BACKGROUND: The complete sequences of chloroplast genomes provide wealthy

information regarding the evolutionary history of species. With the advance of

next-generation sequencing technology, the number of completely sequenced

chloroplast genomes is expected to increase exponentially, powerful computational

tools annotating the genome sequences are in urgent need.

RESULTS: We have developed a web server CPGAVAS. The server accepts a complete

chloroplast genome sequence as input. First, it predicts protein-coding and rRNA

genes based on the identification and mapping of the most similar, full-length

protein, cDNA and rRNA sequences by integrating results from Blastx, Blastn,

protein2genome and est2genome programs. Second, tRNA genes and inverted repeats

(IR) are identified using tRNAscan, ARAGORN and vmatch respectively. Third, it

calculates the summary statistics for the annotated genome. Fourth, it generates

a circular map ready for publication. Fifth, it can create a Sequin file for

GenBank submission. Last, it allows the extractions of protein and mRNA sequences

for given list of genes and species. The annotation results in GFF3 format can be

edited using any compatible annotation editing tools. The edited annotations can

then be uploaded to CPGAVAS for update and re-analyses repeatedly. Using known

chloroplast genome sequences as test set, we show that CPGAVAS performs

comparably to another application DOGMA, while having several superior

functionalities.

CONCLUSIONS: CPGAVAS allows the semi-automatic and complete annotation of a

chloroplast genome sequence, and the visualization, editing and analysis of the

annotation results. It will become an indispensible tool for researchers studying

chloroplast genomes. The software is freely accessible from

http://www.herbalgenomics.org/cpgavas.

DOI: 10.1186/1471-2164-13-715

PMCID: PMC3543216

PMID: 23256920 [Indexed for MEDLINE]

1272. Diagn Pathol. 2012 Dec 13;7:177. doi: 10.1186/1746-1596-7-177.

Development of a teledermatopathology consultation system using virtual slides.

Nakayama I(1), Matsumura T, Kamataki A, Uzuki M, Saito K, Hobbs J, Akasaka T,

Sawai T.

Author information:

(1)Department of Dermatology, Course of Integrated Medicine, Graduate School of

Medicine, Osaka University, Suita, Osaka, Japan.

BACKGROUND: An online consultation system using virtual slides (whole slide

images; WSI) has been developed for pathological diagnosis, and could help

compensate for the shortage of pathologists, especially in the field of

dermatopathology and in other fields dealing with difficult cases. This study

focused on the performance and future potential of the system.

METHOD: In our system, histological specimens on slide glasses are digitalized by

a virtual slide instrument, converted into web data, and up-loaded to an open

server. Using our own purpose-built online system, we then input patient details

such as age, gender, affected region, clinical data, past history and other

related items. We next select up to ten consultants. Finally we send an e-mail to

all consultants simultaneously through a single command. The consultant receives

an e-mail containing an ID and password which is used to access the open server

and inspect the images and other data associated with the case. The consultant

makes a diagnosis, which is sent to us along with comments. Because this was a

pilot study, we also conducted several questionnaires with consultants concerning

the quality of images, operability, usability, and other issues.

RESULTS: We solicited consultations for 36 cases, including cases of tumor, and

involving one to eight consultants in the field of dermatopathology. No problems

were noted concerning the images or the functioning of the system on the sender

or receiver sides. The quickest diagnosis was received only 18 minutes after

sending our data. This is much faster than in conventional consultation using

glass slides. There were no major problems relating to the diagnosis, although

there were some minor differences of opinion between consultants. The results of

questionnaires answered by many consultants confirmed the usability of this

system for pathological consultation. (16 out of 23 consultants.)

CONCLUSION: We have developed a novel teledermatopathological consultation system

using virtual slides, and investigated the usefulness of the system. The results

demonstrate that our system can be a useful tool for international medical work,

and we anticipate its wider application in the future.

VIRTUAL SLIDES: The virtual slides for this article can be found here:

http://www.diagnosticpathology.diagnomx.eu/vs/1902376044831574.

DOI: 10.1186/1746-1596-7-177

PMCID: PMC3557204

PMID: 23237667 [Indexed for MEDLINE]

1273. J Proteome Res. 2012 Dec 7;11(12):6282-90. doi: 10.1021/pr300694b. Epub 2012 Oct

29.

Cloud CPFP: a shotgun proteomics data analysis pipeline using cloud and high

performance computing.

Trudgian DC(1), Mirzaei H.

Author information:

(1)Department of Biochemistry, University of Texas Southwestern Medical Center,

5323 Harry Hines Boulevard, Dallas, Texas 75390-8816, United States.

We have extended the functionality of the Central Proteomics Facilities Pipeline

(CPFP) to allow use of remote cloud and high performance computing (HPC)

resources for shotgun proteomics data processing. CPFP has been modified to

include modular local and remote scheduling for data processing jobs. The

pipeline can now be run on a single PC or server, a local cluster, a remote HPC

cluster, and/or the Amazon Web Services (AWS) cloud. We provide public images

that allow easy deployment of CPFP in its entirety in the AWS cloud. This

significantly reduces the effort necessary to use the software, and allows

proteomics laboratories to pay for compute time ad hoc, rather than obtaining and

maintaining expensive local server clusters. Alternatively the Amazon cloud can

be used to increase the throughput of a local installation of CPFP as necessary.

We demonstrate that cloud CPFP allows users to process data at higher speed than

local installations but with similar cost and lower staff requirements. In

addition to the computational improvements, the web interface to CPFP is

simplified, and other functionalities are enhanced. The software is under active

development at two leading institutions and continues to be released under an

open-source license at http://cpfp.sourceforge.net.

DOI: 10.1021/pr300694b

PMID: 23088505 [Indexed for MEDLINE]

1274. Biochim Biophys Acta. 2012 Dec;1824(12):1425-33. doi:

10.1016/j.bbapap.2012.05.018. Epub 2012 Jun 15.

MetaLocGramN: A meta-predictor of protein subcellular localization for

Gram-negative bacteria.

Magnus M(1), Pawlowski M, Bujnicki JM.

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology, ul. Ks. Trojdena 4, PL-02-109 Warsaw, Poland.

magnus@genesilico.pl

Subcellular localization is a key functional characteristic of proteins. It is

determined by signals encoded in the protein sequence. The experimental

determination of subcellular localization is laborious. Thus, a number of

computational methods have been developed to predict the protein location from

sequence. However predictions made by different methods often disagree with each

other and it is not always clear which algorithm performs best for the given

cellular compartment. We benchmarked primary subcellular localization predictors

for proteins from Gram-negative bacteria, PSORTb3, PSLpred, CELLO, and

SOSUI-GramN, on a common dataset that included 1056 proteins. We found that

PSORTb3 performs best on the average, but is outperformed by other methods in

predictions of extracellular proteins. This motivated us to develop a

meta-predictor, which combines the primary methods by using the logistic

regression models, to take advantage of their combined strengths, and to

eliminate their individual weaknesses. MetaLocGramN runs the primary methods, and

based on their output classifies protein sequences into one of five major

localizations of the Gram-negative bacterial cell: cytoplasm, plasma membrane,

periplasm, outer membrane, and extracellular space. MetaLocGramN achieves the

average Matthews correlation coefficient of 0.806, i.e. 12% better than the best

individual primary method. MetaLocGramN is a meta-predictor specialized in

predicting subcellular localization for proteins from Gram-negative bacteria.

According to our benchmark, it performs better than all other tools run

independently. MetaLocGramN is a web and SOAP server available for free use by

all academic users at the URL http://iimcb.genesilico.pl/MetaLocGramN. This

article is part of a Special Issue entitled: Computational Methods for Protein

Interaction and Structural Prediction.

Copyright © 2012 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.bbapap.2012.05.018

PMID: 22705560 [Indexed for MEDLINE]

1275. Bioinformatics. 2012 Dec 1;28(23):3066-72. doi: 10.1093/bioinformatics/bts598.

Epub 2012 Oct 9.

Predicting protein residue-residue contacts using deep networks and boosting.

Eickholt J(1), Cheng J.

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, MO 65211,

USA.

MOTIVATION: Protein residue-residue contacts continue to play a larger and larger

role in protein tertiary structure modeling and evaluation. Yet, while the

importance of contact information increases, the performance of sequence-based

contact predictors has improved slowly. New approaches and methods are needed to

spur further development and progress in the field.

RESULTS: Here we present DNCON, a new sequence-based residue-residue contact

predictor using deep networks and boosting techniques. Making use of graphical

processing units and CUDA parallel computing technology, we are able to train

large boosted ensembles of residue-residue contact predictors achieving

state-of-the-art performance.

AVAILABILITY: The web server of the prediction method (DNCON) is available at

http://iris.rnet.missouri.edu/dncon/.

CONTACT: chengji@missouri.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/bts598

PMCID: PMC3509494

PMID: 23047561 [Indexed for MEDLINE]

1276. Biopreserv Biobank. 2012 Dec;10(6):501-10. doi: 10.1089/bio.2012.0033.

An online tool for improving biospecimen data element reporting.

Cheah S(1), Dee S, Cole A, Matzke L, O'Donoghue S, Watson PH.

Author information:

(1)1 BC BioLibrary, British Columbia, Canada .

Detailed documentation of the experimental materials and methods is essential for

the validation of scientific papers. Human biospecimens are increasingly utilized

as materials in cancer research and information about the biospecimens used is a

component of this documentation. We hypothesized that previously reported

biospecimen data are inadequate for accurate replication and/or validation of a

substantial proportion of studies. To examine this issue, we analyzed biospecimen

reporting in a representative cross section of publications over the past 12

years (1998, 2004, 2010) in the journals, Cancer Research (CR, n=46) and Clinical

Cancer Research (CCR, n=73). We assessed biospecimen data in relation to the

standards outlined as the Tier 1 recommended data elements from the Biospecimen

Reporting for Improved Study Quality (BRISQ), in addition to ethics criteria.

These data elements encompass features of biospecimens influenced by the patient,

medical procedure, and biospecimen acquisition, handling and storage processes.

Analysis found that while there was a significant increase in the reporting of

ethics board approval status (p<0.008) and name of the ethics board (p<0.0001),

there were no significant differences between these journals or over this period

in reporting other biospecimen-related data elements. Of the 15 Tier 1 data

elements assessed in CR and CCR, the data elements commonly obtained from the

"Clinical Chart" (8/15 elements) were significantly better reported than elements

that would typically be obtained from the "Biobank" (p<0.0001). Our findings

demonstrate that reporting of biospecimen-related data elements has been

incomplete. As one part of the solution to this issue, we propose the use of an

online data-elements reporting tool (www.biobanking.ca) by biobanks. This BRISQ

Report tool aims to help biobanks provide the relevant biospecimen-related data

as a structured report, and to promote its inclusion as supplementary material in

publications to improve the quality of future research studies.

DOI: 10.1089/bio.2012.0033

PMID: 24845136 [Indexed for MEDLINE]

1277. Protein Pept Lett. 2012 Dec;19(12):1318-23.

Defensinpred: defensin and defensin types prediction server.

Kumari SR(1), Badwaik R, Sundararajan V, Jayaraman VK.

Author information:

(1)Centre for Development of Advanced Computing (C-DAC), Pune University Campus,

Pune, India.

Defensins are considered to play an important role in the innate immune system of

virtually all life forms, from insects and plants to amphibians and mammals. They

are classified into alpha, beta and theta-defensins. Fast and accurate

computational prediction of defensin and defensin types will help in annotating

unidentified defensin novel peptides. Identified defensins, owing to their small

length and potent antimicrobial activity can be used effectively for development

of new clinically applicable antibiotics. Thus predicting the defensin candidates

will aid in accurate identification of novel peptide drugs. Support vector

machines prediction model accuracy was 99% for defensin and defensin types. The

results indicate that it is most accurate and efficient prediction method for

defensin peptides. User friendly defensin web server is provided at

www.defensinpred.cdac.in for the benefit of scientific community.

PMID: 22670676 [Indexed for MEDLINE]

1278. Protein Pept Lett. 2012 Dec;19(12):1250-6.

Predicting the metabolic pathways of small molecules based on their

physicochemical properties.

Peng CR(1), Lu WC, Niu B, Li MJ, Yang XY, Wu ML.

Author information:

(1)School of Materials Science and Engineering, Shanghai University, Shanghai

200444, China.

How to correctly and efficiently map small molecule to its possible metabolic

pathway is a meaningful topic to metabonomics research. In this work, a novel

approach to address this problem was introduced to encode physicochemical

properties of small molecules. Based on this encoding method, a two stage feature

selection method called mRMR-FFSAdaBoost was adopted to map small molecules to

their corresponding metabolic pathways possible. As a result, the accuracies of

10-folds cross-validation test and independent set test for predicting the

metabolic pathways of small molecules reached 83.88% and 85.23%, respectively. An

online server for predicting metabolic pathways of unknown small molecules as

described in this paper is accessible at

http://chemdata.shu.edu.cn:8080/PathwayPrediction/.

PMID: 22670666 [Indexed for MEDLINE]

1279. Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi. 2012 Dec;28(12):1324-7.

[Construction of 3D model of CD28 chimeric antibody with its antigen docked].

[Article in Chinese]

Cheng G(1), Qin YY, Cheng D, Chen X, Zhang XG, Qiu YH.

Author information:

(1)Institute of Clinical Immunology, First Affiliated Hospital, Soochow

University, Suzhou 215007, China.

AIM: To construct a 3D model of the chimeric antibodies (AntiCD28: ch-2F5) with

corresponding antigen molecule docked to theoretically verify the rationality of

the binding of antibody with its antigen and to provide a method of 3D

identification between antigen and antibody and spatial structure analysis.

METHODS: We analyzed the sequence by submitting it to

http://www.ncbi.nlm.nih.gov/ and made a comparison using integratly the 3

databases of GenBank, Protein data bank and GENO-3D. The 3D model was constructed

by Swiss-model homology modeling server and molecular docking online was

performed by GRAMM-X Protein Docking Web Server. Chimeric heavy chain, light

chain, heavy-light chain complex, heavy-light chain and antigen complex were

displayed and photographed by the Chimera Software. Meanwhile, the spatial

structures of heavy, light chains, variable region, constant region, CDR and

frame area were marked by different colours respectively to exhibit the 3D

structure on every side.

RESULTS: The 3D structure of the heavy-light chain and antigen complex we

constructed was consistent well with the theory of antigen binding to antibody

molecules.

CONCLUSION: The structure of the chimeric antibody we constructed with the

bioinformatic method was in accordance with the general structure of antibody,

and its antigen binding site was also consistent with the molecular theory. Thus,

the model helps to analyze the 3D structure of antibody and antigen-antibody

interaction.

PMID: 23232527 [Indexed for MEDLINE]

1280. Bioinformatics. 2012 Nov 15;28(22):2998-9. doi: 10.1093/bioinformatics/bts539.

Epub 2012 Sep 6.

SAPIN: a framework for the structural analysis of protein interaction networks.

Yang JS(1), Campagna A, Delgado J, Vanhee P, Serrano L, Kiel C.

Author information:

(1)EMBL/CRG, Design of Biological Systems, Systems Biology Research Unit, Centre

for Genomic Regulation-CRG, UPF, 08003 Barcelona, Spain. jae-seong.yang@crg.eu

SUMMARY: Protein interaction networks are widely used to depict the relationships

between proteins. These networks often lack the information on physical binary

interactions, and they do not inform whether there is incompatibility of

structure between binding partners. Here, we introduce SAPIN, a framework

dedicated to the structural analysis of protein interaction networks. SAPIN first

identifies the protein parts that could be involved in the interaction and

provides template structures. Next, SAPIN performs structural superimpositions to

identify compatible and mutually exclusive interactions. Finally, the results are

displayed using Cytoscape Web.

AVAILABILITY: The SAPIN server is available at http://sapin.crg.es.

CONTACT: jae-seong.yang@crg.eu or christina.kiel@crg.eu.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

Online.

DOI: 10.1093/bioinformatics/bts539

PMID: 22954630 [Indexed for MEDLINE]

1281. IEEE/ACM Trans Comput Biol Bioinform. 2012 Nov-Dec;9(6):1766-75.

Sequence-based prediction of DNA-binding residues in proteins with conservation

and correlation information.

Ma X(1), Guo J, Liu HD, Xie JM, Sun X.

Author information:

(1)State Key Laboratory of Bioelectronics, School of Biological Science & Medical

Engineering, Southeast University and Nanjing Audit University, Nanjing, P.R.

China. maxin@seu.edu.cn

The recognition of DNA-binding residues in proteins is critical to our

understanding of the mechanisms of DNA-protein interactions, gene expression, and

for guiding drug design. Therefore, a prediction method DNABR (DNA Binding

Residues) is proposed for predicting DNA-binding residues in protein sequences

using the random forest (RF) classifier with sequence-based features. Two types

of novel sequence features are proposed in this study, which reflect the

information about the conservation of physicochemical properties of the amino

acids, and the correlation of amino acids between different sequence positions in

terms of physicochemical properties. The first type of feature uses the

evolutionary information combined with the conservation of physicochemical

properties of the amino acids while the second reflects the dependency effect of

amino acids with regards to polarity charge and hydrophobic properties in the

protein sequences. Those two features and an orthogonal binary vector which

reflect the characteristics of 20 types of amino acids are used to build the

DNABR, a model to predict DNA-binding residues in proteins. The DNABR model

achieves a value of 0.6586 for Matthew’s correlation coefficient (MCC) and 93.04

percent overall accuracy (ACC) with a68.47 percent sensitivity (SE) and 98.16

percent specificity (SP), respectively. The comparisons with each feature

demonstrate that these two novel features contribute most to the improvement in

predictive ability. Furthermore, performance comparisons with other approaches

clearly show that DNABR has an excellent prediction performance for detecting

binding residues in putative DNA-binding protein. The DNABR web-server system is

freely available at http://www.cbi.seu.edu.cn/DNABR/.

DOI: 10.1109/TCBB.2012.106

PMID: 22868682 [Indexed for MEDLINE]

1282. J Antimicrob Chemother. 2012 Nov;67(11):2640-4. doi: 10.1093/jac/dks261. Epub

2012 Jul 10.

Identification of acquired antimicrobial resistance genes.

Zankari E(1), Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O,

Aarestrup FM, Larsen MV.

Author information:

(1)National Food Institute, Technical University of Denmark, 2800 Kgs. Lyngby,

Denmark. east@food.dtu.dk

OBJECTIVES: Identification of antimicrobial resistance genes is important for

understanding the underlying mechanisms and the epidemiology of antimicrobial

resistance. As the costs of whole-genome sequencing (WGS) continue to decline, it

becomes increasingly available in routine diagnostic laboratories and is

anticipated to substitute traditional methods for resistance gene identification.

Thus, the current challenge is to extract the relevant information from the large

amount of generated data.

METHODS: We developed a web-based method, ResFinder that uses BLAST for

identification of acquired antimicrobial resistance genes in whole-genome data.

As input, the method can use both pre-assembled, complete or partial genomes, and

short sequence reads from four different sequencing platforms. The method was

evaluated on 1862 GenBank files containing 1411 different resistance genes, as

well as on 23 de-novo-sequenced isolates.

RESULTS: When testing the 1862 GenBank files, the method identified the

resistance genes with an ID = 100% (100% identity) to the genes in ResFinder.

Agreement between in silico predictions and phenotypic testing was found when the

method was further tested on 23 isolates of five different bacterial species,

with available phenotypes. Furthermore, ResFinder was evaluated on WGS

chromosomes and plasmids of 30 isolates. Seven of these isolates were annotated

to have antimicrobial resistance, and in all cases, annotations were compatible

with the ResFinder results.

CONCLUSIONS: A web server providing a convenient way of identifying acquired

antimicrobial resistance genes in completely sequenced isolates was created.

ResFinder can be accessed at www.genomicepidemiology.org. ResFinder will

continuously be updated as new resistance genes are identified.

DOI: 10.1093/jac/dks261

PMCID: PMC3468078

PMID: 22782487 [Indexed for MEDLINE]

1283. Med Chem. 2012 Nov;8(6):1108-16.

QSAR study on 5-lipoxygenase inhibitors based on support vector machine.

Niu B(1), Su Q, Yuan X, Lu W, Ding J.

Author information:

(1)College of Life Science, Shanghai University, Shanghai 200444, China.

bingniu@shu.edu.cn

QSAR study on a data set of 5-lipoxygenase inhibitors (1-phenyl

[2H]-tetrahydro-triazine-3-one analogues) was carried out by using Support Vector

Regression (SVR) and physicochemical parameters. Wrapper methods were used to

select descriptors, while Leave-One-Out Cross Validation (LOOCV) method and

independent set test were used to judge the predictive power of different models.

We found out that the generalization ability of SVR model outperformed multiple

linear regression (MLR) and Partial Least Squares (PLS) models in this work. An

online web server for activity prediction is available at

http://chemdata.shu.edu.cn/qsar5lip.

PMID: 22779798 [Indexed for MEDLINE]

1284. Mol Biosyst. 2012 Nov;8(11):2964-73. doi: 10.1039/c2mb25251a.

Systematic analysis of human lysine acetylation proteins and accurate prediction

of human lysine acetylation through bi-relative adapted binomial score Bayes

feature representation.

Shao J(1), Xu D, Hu L, Kwan YW, Wang Y, Kong X, Ngai SM.

Author information:

(1)Institute of Health Sciences, Shanghai Institutes of Biological Sciences,

Chinese Academy of Sciences, Shanghai, People's Republic of China.

Lysine acetylation is a reversible post-translational modification (PTM) which

has been linked to many biological and pathological implications. Hence,

localization of lysine acetylation is essential for deciphering the mechanism of

such implications. Whereas many acetylated lysines in human proteins have been

localized through experimental approaches in wet lab, it still fails to reach

completion. In the present study, we proposed a novel feature extraction

approach, bi-relative adapted binomial score Bayes (BRABSB), combined with

support vector machines (SVMs) to construct a human-specific lysine acetylation

predictor, which yields, on average, a sensitivity of 83.91%, a specificity of

87.25% and an accuracy of 85.58%, in the case of 5-fold cross validation

experiments. Results obtained through the validation on independent data sets

show that the proposed approach here outperforms other existing lysine

acetylation predictors. Furthermore, due to the fact that global analysis of

human lysine acetylproteins, which would ultimately facilitate the systematic

investigation of the biological and pathological consequences associated with

lysine acetylation events, remains to be resolved, we made an attempt to

systematically analyze human lysine acetylproteins, demonstrating their diversity

with respect to subcellular localization as well as biological process and

predominance by "binding" in terms of molecular function. Our analysis also

revealed that human lysine acetylproteins are significantly enriched in

neurodegenerative disorders and cancer pathways. Remarkably, lysine

acetylproteins in mitochondria are significantly related to neurodegenerative

disorders and those in the nucleus are instead significantly involved in pathways

in cancers, all of which might ultimately provide novel global insights into such

pathological processes for the therapeutic purpose. The web server is deployed at

http://www.bioinfo.bio.cuhk.edu.hk/bpbphka.

DOI: 10.1039/c2mb25251a

PMID: 22936054 [Indexed for MEDLINE]

1285. Protein Eng Des Sel. 2012 Nov;25(11):705-13. doi: 10.1093/protein/gzs081. Epub

2012 Oct 16.

Reliable and robust detection of coevolving protein residues.

Jeong CS(1), Kim D.

Author information:

(1)Department of Bio and Brain Engineering, KAIST, Daejeon 305-701, South Korea.

Since the cooperative mechanism between interconnected residues plays a critical

role in protein functions, the detection of coevolving residues is important for

studying various biological functions of proteins. In this work, we developed a

new correlated mutation analysis method that shows substantially better

prediction accuracy than all other methods. More importantly, the prediction

accuracy of our new method is insensitive to the characteristics of the multiple

sequence alignments (MSAs) from which the correlated mutation scores are

calculated. Thanks to this desirable property, not only it does it show a good

performance even for MSAs automatically generated by sequence homology

methodologies, which allows us to build a fully automatic easy-to-use server

named CMAT, but its performance is also consistently high on the columns of MSAs

containing a high fraction of gaps, which greatly extends the applicability of

the correlated mutation analysis. The key development of this work is the joint

probability estimation that can be greatly improved by utilizing sequence profile

as prior knowledge, which is shown to be highly beneficial to the correlated

mutation analysis and its applications. From the computational perspective, we

made two important findings; the sequence profile can be used to estimate the

pseudocounts, and the consistency rule on joint probabilities and marginal

probabilities is important for accurately estimating the joint probability. The

web server and standalone program are freely available on the web at

http://binfolab12.kaist.ac.kr/cmat/.

DOI: 10.1093/protein/gzs081

PMID: 23077274 [Indexed for MEDLINE]

1286. Bioinformatics. 2012 Oct 15;28(20):2608-14. doi: 10.1093/bioinformatics/bts493.

Application of asymmetric statistical potentials to antibody-protein docking.

Brenke R(1), Hall DR, Chuang GY, Comeau SR, Bohnuud T, Beglov D, Schueler-Furman

O, Vajda S, Kozakov D.

Author information:

(1)Department of Biomedical Engineering, Boston University, Boston, MA, USA.

MOTIVATION: An effective docking algorithm for antibody-protein antigen complex

prediction is an important first step toward design of biologics and vaccines. We

have recently developed a new class of knowledge-based interaction potentials

called Decoys as the Reference State (DARS) and incorporated DARS into the

docking program PIPER based on the fast Fourier transform correlation approach.

Although PIPER was the best performer in the latest rounds of the CAPRI protein

docking experiment, it is much less accurate for docking antibody-protein antigen

pairs than other types of complexes, in spite of incorporating sequence-based

information on the location of the paratope. Analysis of antibody-protein antigen

complexes has revealed an inherent asymmetry within these interfaces.

Specifically, phenylalanine, tryptophan and tyrosine residues highly populate the

paratope of the antibody but not the epitope of the antigen.

RESULTS: Since this asymmetry cannot be adequately modeled using a symmetric

pairwise potential, we have removed the usual assumption of symmetry. Interaction

statistics were extracted from antibody-protein complexes under the assumption

that a particular atom on the antibody is different from the same atom on the

antigen protein. The use of the new potential significantly improves the

performance of docking for antibody-protein antigen complexes, even without any

sequence information on the location of the paratope. We note that the asymmetric

potential captures the effects of the multi-body interactions inherent to the

complex environment in the antibody-protein antigen interface.

AVAILABILITY: The method is implemented in the ClusPro protein docking server,

available at http://cluspro.bu.edu.

DOI: 10.1093/bioinformatics/bts493

PMCID: PMC3467743

PMID: 23053206 [Indexed for MEDLINE]

1287. Bioinformatics. 2012 Oct 15;28(20):2696-7. doi: 10.1093/bioinformatics/bts506.

Epub 2012 Aug 24.

TRFolder-W: a web server for telomerase RNA structure prediction in yeast

genomes.

Zhang D(1), Xue X, Malmberg RL, Cai L.

Author information:

(1)Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA.

TRFolder-W is a web server capable of predicting core structures of telomerase

RNA (TR) in yeast genomes. TRFolder is a command-line Python toolkit for

TR-specific structure prediction. We developed a web-version built on the django

web framework, leveraging the work done previously, to include enhancements to

increase flexibility of usage. To date, there are five core sub-structures

commonly found in TR of fungal species, which are the template region, downstream

pseudoknot, boundary element, core-closing stem and triple helix. The aim of

TRFolder-W is to use the five core structures as fundamental units to predict

potential TR genes for yeast, and to provide a user-friendly interface. Moreover,

the application of TRFolder-W can be extended to predict the characteristic

structure on species other than fungal species.AVAILABILITY: The web server

TRFolder-W is available at

http://rna-informatics.uga.edu/?f=software&p=TRFolder-w.

DOI: 10.1093/bioinformatics/bts506

PMCID: PMC3467749

PMID: 22923293 [Indexed for MEDLINE]

1288. Bioinformatics. 2012 Oct 15;28(20):2687-8. doi: 10.1093/bioinformatics/bts510.

Epub 2012 Aug 24.

EFICAz2.5: application of a high-precision enzyme function predictor to 396

proteomes.

Kumar N(1), Skolnick J.

Author information:

(1)Center for Study of Systems Biology, School of Biology, Georgia Institute of

Technology, Atlanta, GA 30318, USA.

High-quality enzyme function annotation is essential for understanding the

biochemistry, metabolism and disease processes of organisms. Previously, we

developed a multi-component high-precision enzyme function predictor, EFICAz(2)

(enzyme function inference by a combined approach). Here, we present an updated

improved version, EFICAz(2.5), that is trained on a significantly larger data set

of enzyme sequences and PROSITE patterns. We also present the results of the

application of EFICAz(2.5) to the enzyme reannotation of 396 genomes cataloged in

the ENSEMBL database.AVAILABILITY: The EFICAz(2.5) server and database is freely

available with a use-friendly interface at

http://cssb.biology.gatech.edu/EFICAz2.5.

DOI: 10.1093/bioinformatics/bts510

PMCID: PMC3467752

PMID: 22923291 [Indexed for MEDLINE]

1289. J Chem Theory Comput. 2012 Oct 9;8(10):3618-27. doi: 10.1021/ct3000662. Epub 2012

Jun 6.

SQUEEZE-E: The Optimal Solution for Molecular Simulations with Periodic Boundary

Conditions.

Wassenaar TA(1,)(2), de Vries S(2), Bonvin AM(2), Bekker H(3).

Author information:

(1)Molecular Dynamics Group, Groningen Institute for Biotechnology and

Biomolecular Sciences, University of Groningen, Nijenborgh 7, 9747 AG, Groningen,

The Netherlands. (2)Bijvoet Center for Biomolecular Research, Faculty of Science,

Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands. (3)Johann

Bernoulli Institute for Mathematics and Computer Science, University of

Groningen, POB 800, 9700 AV, Groningen, The Netherlands.

In molecular simulations of macromolecules, it is desirable to limit the amount

of solvent in the system to avoid spending computational resources on

uninteresting solvent-solvent interactions. As a consequence, periodic boundary

conditions are commonly used, with a simulation box chosen as small as possible,

for a given minimal distance between images. Here, we describe how such a

simulation cell can be set up for ensembles, taking into account a priori

available or estimable information regarding conformational flexibility. Doing so

ensures that any conformation present in the input ensemble will satisfy the

distance criterion during the simulation. This helps avoid periodicity artifacts

due to conformational changes. The method introduces three new approaches in

computational geometry: (1) The first is the derivation of an optimal packing of

ensembles, for which the mathematical framework is described. (2) A new method

for approximating the α-hull and the contact body for single bodies and ensembles

is presented, which is orders of magnitude faster than existing routines,

allowing the calculation of packings of large ensembles and/or large bodies. 3. A

routine is described for searching a combination of three vectors on a

discretized contact body forming a reduced base for a lattice with minimal cell

volume. The new algorithms reduce the time required to calculate packings of

single bodies from minutes or hours to seconds. The use and efficacy of the

method is demonstrated for ensembles obtained from NMR, MD simulations, and

elastic network modeling. An implementation of the method has been made available

online at http://haddock.chem.uu.nl/services/SQUEEZE/ and has been made available

as an option for running simulations through the weNMR GRID MD server at

http://haddock.science.uu.nl/enmr/services/GROMACS/main.php .

DOI: 10.1021/ct3000662

PMID: 26593007

1290. BMC Genomics. 2012 Oct 6;13:535. doi: 10.1186/1471-2164-13-535.

GeneFriends: an online co-expression analysis tool to identify novel gene targets

for aging and complex diseases.

van Dam S(1), Cordeiro R, Craig T, van Dam J, Wood SH, de Magalhães JP.

Author information:

(1)Integrative Genomics of Ageing Group, Institute of Integrative Biology,

University of Liverpool, Liverpool, L69 7ZB, UK.

BACKGROUND: Although many diseases have been well characterized at the molecular

level, the underlying mechanisms are often unknown. Nearly half of all human

genes remain poorly studied, yet these genes may contribute to a number of

disease processes. Genes involved in common biological processes and diseases are

often co-expressed. Using known disease-associated genes in a co-expression

analysis may help identify and prioritize novel candidate genes for further

study.

RESULTS: We have created an online tool, called GeneFriends, which identifies

co-expressed genes in over 1,000 mouse microarray datasets. GeneFriends can be

used to assign putative functions to poorly studied genes. Using a seed list of

disease-associated genes and a guilt-by-association method, GeneFriends allows

users to quickly identify novel genes and transcription factors associated with a

disease or process. We tested GeneFriends using seed lists for aging, cancer, and

mitochondrial complex I disease. We identified several candidate genes that have

previously been predicted as relevant targets. Some of the genes identified are

already being tested in clinical trials, indicating the effectiveness of this

approach. Co-expressed transcription factors were investigated, identifying C/ebp

genes as candidate regulators of aging. Furthermore, several novel candidate

genes, that may be suitable for experimental or clinical follow-up, were

identified. Two of the novel candidates of unknown function that were

co-expressed with cancer-associated genes were selected for experimental

validation. Knock-down of their human homologs (C1ORF112 and C12ORF48) in HeLa

cells slowed growth, indicating that these genes of unknown function, identified

by GeneFriends, may be involved in cancer.

CONCLUSIONS: GeneFriends is a resource for biologists to identify and prioritize

novel candidate genes involved in biological processes and complex diseases. It

is an intuitive online resource that will help drive experimentation. GeneFriends

is available online at: http://genefriends.org/.

DOI: 10.1186/1471-2164-13-535

PMCID: PMC3495651

PMID: 23039964 [Indexed for MEDLINE]

1291. Front Genet. 2012 Oct 5;3:197. doi: 10.3389/fgene.2012.00197. eCollection 2012.

The ontology-based answers (OBA) service: a connector for embedded usage of

ontologies in applications.

Dönitz J(1), Wingender E.

Author information:

(1)Department of Bioinformatics, University Medical Center Göttingen Göttingen,

Germany ; Department of Developmental Biology, Johann-Friedrich-Blumenbach

Institute for Zoology and Anthropology, Georg-August University Göttingen

Göttingen, Germany.

The semantic web depends on the use of ontologies to let electronic systems

interpret contextual information. Optimally, the handling and access of

ontologies should be completely transparent to the user. As a means to this end,

we have developed a service that attempts to bridge the gap between experts in a

certain knowledge domain, ontologists, and application developers. The

ontology-based answers (OBA) service introduced here can be embedded into custom

applications to grant access to the classes of ontologies and their relations as

most important structural features as well as to information encoded in the

relations between ontology classes. Thus computational biologists can benefit

from ontologies without detailed knowledge about the respective ontology. The

content of ontologies is mapped to a graph of connected objects which is

compatible to the object-oriented programming style in Java. Semantic functions

implement knowledge about the complex semantics of an ontology beyond the class

hierarchy and "partOf" relations. By using these OBA functions an application

can, for example, provide a semantic search function, or (in the examples

outlined) map an anatomical structure to the organs it belongs to. The semantic

functions relieve the application developer from the necessity of acquiring

in-depth knowledge about the semantics and curation guidelines of the used

ontologies by implementing the required knowledge. The architecture of the OBA

service encapsulates the logic to process ontologies in order to achieve a

separation from the application logic. A public server with the current plugins

is available and can be used with the provided connector in a custom application

in scenarios analogous to the presented use cases. The server and the client are

freely available if a project requires the use of custom plugins or non-public

ontologies. The OBA service and further documentation is available at

http://www.bioinf.med.uni-goettingen.de/projects/oba.

DOI: 10.3389/fgene.2012.00197

PMCID: PMC3464866

PMID: 23060901

1292. BMC Biol. 2012 Oct 2;10:82. doi: 10.1186/1741-7007-10-82.

MESSA: MEta-Server for protein Sequence Analysis.

Cong Q(1), Grishin NV.

Author information:

(1)Department of Biophysics, University of Texas Southwestern Medical Center,

Dallas, TX 75390-9050, USA.

BACKGROUND: Computational sequence analysis, that is, prediction of local

sequence properties, homologs, spatial structure and function from the sequence

of a protein, offers an efficient way to obtain needed information about proteins

under study. Since reliable prediction is usually based on the consensus of many

computer programs, meta-severs have been developed to fit such needs. Most

meta-servers focus on one aspect of sequence analysis, while others incorporate

more information, such as PredictProtein for local sequence feature predictions,

SMART for domain architecture and sequence motif annotation, and GeneSilico for

secondary and spatial structure prediction. However, as predictions of local

sequence properties, three-dimensional structure and function are usually

intertwined, it is beneficial to address them together.

RESULTS: We developed a MEta-Server for protein Sequence Analysis (MESSA) to

facilitate comprehensive protein sequence analysis and gather structural and

functional predictions for a protein of interest. For an input sequence, the

server exploits a number of select tools to predict local sequence properties,

such as secondary structure, structurally disordered regions, coiled coils,

signal peptides and transmembrane helices; detect homologous proteins and assign

the query to a protein family; identify three-dimensional structure templates and

generate structure models; and provide predictive statements about the protein's

function, including functional annotations, Gene Ontology terms, enzyme

classification and possible functionally associated proteins. We tested MESSA on

the proteome of Candidatus Liberibacter asiaticus. Manual curation shows that

three-dimensional structure models generated by MESSA covered around 75% of all

the residues in this proteome and the function of 80% of all proteins could be

predicted.

AVAILABILITY: MESSA is free for non-commercial use at

http://prodata.swmed.edu/MESSA/

DOI: 10.1186/1741-7007-10-82

PMCID: PMC3519821

PMID: 23031578 [Indexed for MEDLINE]

1293. Bioinformatics. 2012 Oct 1;28(19):2517-9. Epub 2012 Jul 24.

Nebula--a web-server for advanced ChIP-seq data analysis.

Boeva V(1), Lermine A, Barette C, Guillouf C, Barillot E.

Author information:

(1)Institut Curie, Paris 75248, France.

MOTIVATION: ChIP-seq consists of chromatin immunoprecipitation and deep

sequencing of the extracted DNA fragments. It is the technique of choice for

accurate characterization of the binding sites of transcription factors and other

DNA-associated proteins. We present a web service, Nebula, which allows

inexperienced users to perform a complete bioinformatics analysis of ChIP-seq

data.

RESULTS: Nebula was designed for both bioinformaticians and biologists. It is

based on the Galaxy open source framework. Galaxy already includes a large number

of functionalities for mapping reads and peak calling. We added the following to

Galaxy: (i) peak calling with FindPeaks and a module for immunoprecipitation

quality control, (ii) de novo motif discovery with ChIPMunk, (iii) calculation of

the density and the cumulative distribution of peak locations relative to gene

transcription start sites, (iv) annotation of peaks with genomic features and (v)

annotation of genes with peak information. Nebula generates the graphs and the

enrichment statistics at each step of the process. During Steps 3-5, Nebula

optionally repeats the analysis on a control dataset and compares these results

with those from the main dataset. Nebula can also incorporate gene expression (or

gene modulation) data during these steps. In summary, Nebula is an innovative web

service that provides an advanced ChIP-seq analysis pipeline providing

ready-to-publish results.

AVAILABILITY: Nebula is available at http://nebula.curie.fr/

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/bts463

PMID: 22829625 [Indexed for MEDLINE]

1294. Bioinformatics. 2012 Oct 1;28(19):2523-6. Epub 2012 Jul 23.

RILogo: visualizing RNA-RNA interactions.

Menzel P(1), Seemann SE, Gorodkin J.

Author information:

(1)The Bioinformatics Centre, Department of Biology, University of Copenhagen,

Ole Maaløes Vej 5, Copenhagen DK-2200, Denmark. pmenzel@gmail.com

SUMMARY: With the increasing amount of newly discovered non-coding RNAs, the

interactions between RNA molecules become an increasingly important aspect for

characterizing their functionality. Many computational tools have been developed

to predict the formation of duplexes between two RNAs, either based on single

sequences or alignments of homologous sequences. Here, we present RILogo, a

program to visualize inter- and intramolecular base pairing between two RNA

molecules. The input for RILogo is a pair of structure-annotated sequences or

alignments. In the latter case, RILogo displays the alignments in the form of

sequence logos, including the mutual information of base paired columns. We also

introduce two novel mutual information based measures that weigh the covariance

information by the evolutionary distances of the aligned sequences. We show that

the new measures have an increased accuracy compared with previous mutual

information measures.

AVAILABILITY AND IMPLEMENTATION: RILogo is freely available as a stand-alone

program and is accessible via a web server at http://rth.dk/resources/rilogo.

CONTACT: pmenzel@gmail.com

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/bts461

PMID: 22826541 [Indexed for MEDLINE]

1295. Bioinformatics. 2012 Oct 1;28(19):2509-11. Epub 2012 Jul 18.

UniMoG--a unifying framework for genomic distance calculation and sorting based

on DCJ.

Hilker R(1), Sickinger C, Pedersen CN, Stoye J.

Author information:

(1)Computational Genomics, Bielefeld University, 33615 Bielefeld, Germany.

rhilker@cebitec.uni-bielefeld.de

SUMMARY: UniMoG is a software combining five genome rearrangement models: double

cut and join (DCJ), restricted DCJ, Hannenhalli and Pevzner (HP), inversion and

translocation. It can compute the pairwise genomic distances and a corresponding

optimal sorting scenario for an arbitrary number of genomes. All five models can

be unified through the DCJ model, thus the implementation is based on DCJ and,

where reasonable, uses the most efficient existing algorithms for each distance

and sorting problem. Both textual and graphical output is possible for

visualizing the operations.

AVAILABILITY AND IMPLEMENTATION: The software is available through the Bielefeld

University Bioinformatics Web Server at

http://bibiserv.techfak.uni-bielefeld.de/dcj with instructions and example data.

CONTACT: rhilker@cebitec.uni-bielefeld.de.

DOI: 10.1093/bioinformatics/bts440

PMCID: PMC3463123

PMID: 22815356 [Indexed for MEDLINE]

1296. Future Med Chem. 2012 Oct;4(15):1933-44. doi: 10.4155/fmc.12.152.

Computational tools and resources for metabolism-related property predictions. 2.

Application to prediction of half-life time in human liver microsomes.

Zakharov AV(1), Peach ML, Sitzmann M, Filippov IV, McCartney HJ, Smith LH,

Pugliese A, Nicklaus MC.

Author information:

(1)CADD Group, Chemical Biology Laboratory, Center for Cancer Research, National

Cancer Institute, National Institutes of Health, DHHS, Frederick National

Laboratory for Cancer Research, Frederick, MD 21702, USA.

BACKGROUND: The most important factor affecting metabolic excretion of compounds

from the body is their half-life time. This provides an indication of compound

stability of, for example, drug molecules. We report on our efforts to develop

QSAR models for metabolic stability of compounds, based on in vitro half-life

assay data measured in human liver microsomes.

METHOD: A variety of QSAR models generated using different statistical methods

and descriptor sets implemented in both open-source and commercial programs

(KNIME, GUSAR and StarDrop) were analyzed. The models obtained were compared

using four different external validation sets from public and commercial data

sources, including two smaller sets of in vivo half-life data in humans.

CONCLUSION: In many cases, the accuracy of prediction achieved on one external

test set did not correspond to the results achieved with another test set. The

most predictive models were used for predicting the metabolic stability of

compounds from the open NCI database, the results of which are publicly available

on the NCI/CADD Group web server ( http://cactus.nci.nih.gov ).

DOI: 10.4155/fmc.12.152

PMCID: PMC4117347

PMID: 23088274 [Indexed for MEDLINE]

1297. Genomics Proteomics Bioinformatics. 2012 Oct;10(5):310-6. doi:

10.1016/j.gpb.2012.08.005. Epub 2012 Sep 29.

miRT: a database of validated transcription start sites of human microRNAs.

Bhattacharyya M(1), Das M, Bandyopadhyay S.

Author information:

(1)Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108,

India.

MicroRNAs (miRNAs) are small endogenous non-coding RNAs of about 22 nt in length

that take crucial roles in many biological processes. These short RNAs regulate

the expression of mRNAs by binding to their 3'-UTRs or by translational

repression. Many of the current studies focus on how mature miRNAs regulate

mRNAs, however, very limited knowledge is available regarding their

transcriptional loci. It is known that primary miRNAs (pri-miRs) are first

transcribed from the DNA, followed by the formation of precursor miRNAs

(pre-miRs) by endonuclease activity, which finally produces the mature miRNAs.

Till date, many of the pre-miRs and mature miRNAs have been experimentally

verified. But unfortunately, identification of the loci of pri-miRs, promoters

and associated transcription start sites (TSSs) are still in progress. TSSs of

only about 40% of the known mature miRNAs in human have been reported. This

information, albeit limited, may be useful for further study of the regulation of

miRNAs. In this paper, we provide a novel database of validated miRNA TSSs, miRT,

by collecting data from several experimental studies that validate miRNA TSSs and

are available for full download. We present miRT as a web server and it is also

possible to convert the TSS loci between different genome built. miRT might be a

valuable resource for advanced research on miRNA regulation, which is freely

accessible at: http://www.isical.ac.in/~bioinfo\_miu/miRT/miRT.php.

Copyright © 2012. Published by Elsevier Ltd.

DOI: 10.1016/j.gpb.2012.08.005

PMCID: PMC5054196

PMID: 23200141 [Indexed for MEDLINE]

1298. J Biomol NMR. 2012 Oct;54(2):169-79. Epub 2012 Aug 19.

PACSY, a relational database management system for protein structure and chemical

shift analysis.

Lee W(1), Yu W, Kim S, Chang I, Lee W, Markley JL.

Author information:

(1)National Magnetic Resonance Facility at Madison, and Biochemistry Department,

University of Wisconsin-Madison, Madison, WI 53706, USA. whlee@nmrfam.wisc.edu

PACSY (Protein structure And Chemical Shift NMR spectroscopY) is a relational

database management system that integrates information from the Protein Data

Bank, the Biological Magnetic Resonance Data Bank, and the Structural

Classification of Proteins database. PACSY provides three-dimensional coordinates

and chemical shifts of atoms along with derived information such as torsion

angles, solvent accessible surface areas, and hydrophobicity scales. PACSY

consists of six relational table types linked to one another for coherence by key

identification numbers. Database queries are enabled by advanced search functions

supported by an RDBMS server such as MySQL or PostgreSQL. PACSY enables users to

search for combinations of information from different database sources in support

of their research. Two software packages, PACSY Maker for database creation and

PACSY Analyzer for database analysis, are available from

http://pacsy.nmrfam.wisc.edu.

DOI: 10.1007/s10858-012-9660-3

PMCID: PMC3542970

PMID: 22903636 [Indexed for MEDLINE]

1299. J Struct Biol. 2012 Oct;180(1):226-34. doi: 10.1016/j.jsb.2012.05.011. Epub 2012

Jun 1.

Determining pair distance distribution function from SAXS data using parametric

functionals.

Liu H(1), Zwart PH.

Author information:

(1)Physical Biosciences Division, Lawrence Berkeley National Laboratories, One

Cyclotron Road, Berkeley, CA 94720, USA. haiguang.liu@asu.edu

Small angle X-ray scattering (SAXS) experiments are widely applied in structural

biology. The SAXS experiments yield one-dimensional profile that needs further

analysis to reveal structural information. The pair distance distribution

function (PDDF), P(r), can provide molecular structures more intuitively, and it

can be used to guide ab initio model reconstructions, making it a critical step

to derive P(r) from experimental SAXS profiles. To calculate the P(r) curves, a

new method based on a specially designed parametric functional form is developed,

and implemented in pregxs. This method is tested against both synthetic and

experimental data, the estimated P(r) functions are in good agreement with

correct or known P(r). The method can also predict the molecular size. In

summary, the pregxs method is robust and accurate in P(r) determination from SAXS

profiles. The pregxs source code and an online server are available at

http://www.sastbx.als.lbl.gov.

Published by Elsevier Inc.

DOI: 10.1016/j.jsb.2012.05.011

PMID: 22659403 [Indexed for MEDLINE]

1300. BMC Res Notes. 2012 Sep 25;5:530. doi: 10.1186/1756-0500-5-530.

Structural attributes for the recognition of weak and anomalous regions in

coiled-coils of myosins and other motor proteins.

Sunitha MS(1), Nair AG, Charya A, Jadhav K, Mukhopadhyay S, Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences (TIFR), UAS-GKVK Campus, Bellary Road,

Bangalore 560 065, India.

BACKGROUND: Coiled-coils are found in different proteins like transcription

factors, myosin tail domain, tropomyosin, leucine zippers and kinesins. Analysis

of various structures containing coiled-coils has revealed the importance of

electrostatic and hydrophobic interactions. In such domains, regions of different

strength of interactions need to be identified since they could be biologically

relevant.

FINDINGS: We have updated our coiled-coil validation webserver, now called

COILCHECK+, where new features were added to efficiently identify the strength of

interaction at the interface region and measure the density of charged residues

and hydrophobic residues. We have examined charged residues and hydrophobic

ladders, using a new algorithm called CHAHO, which is incorporated within

COILCHECK + server. CHAHO permits the identification of spatial charged residue

patches and the continuity of hydrophobic ladder which stabilizes and

destabilizes the coiled-coil structure.

CONCLUSIONS: The availability of such computational tools should be useful to

understand the importance of spatial clustering of charged residues and the

continuity of hydrophobic residues at the interface region of coiled-coil dimers.

COILCHECK + is a structure based tool to validate coiled-coil stability; it can

be accessed at http://caps.ncbs.res.in/coilcheckplus.

DOI: 10.1186/1756-0500-5-530

PMCID: PMC3542152

PMID: 23009691 [Indexed for MEDLINE]

1301. Bioinformatics. 2012 Sep 15;28(18):2391-3. doi: 10.1093/bioinformatics/bts446.

Epub 2012 Jul 12.

RIBFIND: a web server for identifying rigid bodies in protein structures and to

aid flexible fitting into cryo EM maps.

Pandurangan AP(1), Topf M.

Author information:

(1)Department of Crystallography/Biological Sciences, Institute of Structural and

Molecular Biology, Birkbeck College, University of London, London WC1E 7HX, UK.

a.pandurangan@mail.cryst.bbk.ac.uk

MOTIVATION: To better analyze low-resolution cryo electron microscopy maps of

macromolecular assemblies, component atomic structures frequently have to be

flexibly fitted into them. Reaching an optimal fit and preventing the fitting

process from getting trapped in local minima can be significantly improved by

identifying appropriate rigid bodies (RBs) in the fitted component.

RESULTS: Here we present the RIBFIND server, a tool for identifying RBs in

protein structures. The server identifies RBs in proteins by calculating spatial

proximity between their secondary structural elements.

AVAILABILITY: The RIBFIND web server and its standalone program are available at

http://ribfind.ismb.lon.ac.uk.

CONTACT: a.pandurangan@mail.cryst.bbk.ac.uk

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/bts446

PMID: 22796953 [Indexed for MEDLINE]

1302. BMC Bioinformatics. 2012 Sep 11;13:225. doi: 10.1186/1471-2105-13-225.

Coupled mutation finder: a new entropy-based method quantifying phylogenetic

noise for the detection of compensatory mutations.

Gültas M(1), Haubrock M, Tüysüz N, Waack S.

Author information:

(1)Institute of Computer Science, University of Göttingen, Goldschmidtstr. 7,

Göttingen, 37077, Germany. gueltas@cs.uni-goettingen.de

BACKGROUND: The detection of significant compensatory mutation signals in

multiple sequence alignments (MSAs) is often complicated by noise. A challenging

problem in bioinformatics is remains the separation of significant signals

between two or more non-conserved residue sites from the phylogenetic noise and

unrelated pair signals. Determination of these non-conserved residue sites is as

important as the recognition of strictly conserved positions for understanding of

the structural basis of protein functions and identification of functionally

important residue regions. In this study, we developed a new method, the Coupled

Mutation Finder (CMF) quantifying the phylogenetic noise for the detection of

compensatory mutations.

RESULTS: To demonstrate the effectiveness of this method, we analyzed essential

sites of two human proteins: epidermal growth factor receptor (EGFR) and

glucokinase (GCK). Our results suggest that the CMF is able to separate

significant compensatory mutation signals from the phylogenetic noise and

unrelated pair signals. The vast majority of compensatory mutation sites found by

the CMF are related to essential sites of both proteins and they are likely to

affect protein stability or functionality.

CONCLUSIONS: The CMF is a new method, which includes an MSA-specific statistical

model based on multiple testing procedures that quantify the error made in terms

of the false discovery rate and a novel entropy-based metric to upscale BLOSUM62

dissimilar compensatory mutations. Therefore, it is a helpful tool to predict and

investigate compensatory mutation sites of structural or functional importance in

proteins. We suggest that the CMF could be used as a novel automated function

prediction tool that is required for a better understanding of the structural

basis of proteins. The CMF server is freely accessible at

http://cmf.bioinf.med.uni-goettingen.de.

DOI: 10.1186/1471-2105-13-225

PMCID: PMC3577461

PMID: 22963049 [Indexed for MEDLINE]

1303. Neurology. 2012 Sep 11;79(11):1084-93. Epub 2012 Aug 15.

Development of an online tool to determine appropriateness for an epilepsy

surgery evaluation.

Jette N(1), Quan H, Tellez-Zenteno JF, Macrodimitris S, Hader WJ, Sherman EM,

Hamiwka LD, Wirrell EC, Burneo JG, Metcalfe A, Faris PD, Hernandez-Ronquillo L,

Kwon CS, Kirk A, Wiebe S; CASES Expert Panelists.

Collaborators: Andermann F, Burneo JG, Camfield P, Carmant L, Davenport WJ,

Farmer JP, Gross DW, Hader WJ, Huntsman R, Sadler RM, Snead CO 3rd, Steven DA,

Wheatley M.

Author information:

(1)Department of Clinical Neurosciences and Hotchkiss BrainInstitute, University

of Calgary, Calgary, Alberta, Canada. nathalie.jette@albertahealthservices.ca

Comment in

Neurology. 2013 Jun 4;80(23):2169.

Neurology. 2013 Jun 4;80(23):2169.

Neurology. 2012 Sep 11;79(11):1074-5.

OBJECTIVES: Despite evidence that epilepsy surgery is more effective than medical

therapy, significant delays between seizure intractability and surgery exist. We

aimed to develop a new Web-based methodology to assist physicians in identifying

patients who might benefit from an epilepsy surgery evaluation.

METHODS: The RAND/UCLA appropriateness method was used. Clinical scenarios were

developed based on eligibility criteria from previously published surgical

series. Thirteen national experts rated the scenarios for their appropriateness

for an epilepsy surgery evaluation based on published evidence. All scenarios

were rerated after a face-to-face meeting following a modified Delphi process.

Appropriate scenarios were rerated for necessity to determine referral priority.

RESULTS: Of the final 2646 scenarios, 20.6% (n = 544) were appropriate, 17.2% (n

= 456) uncertain, and 61.5% (n = 1626) inappropriate for a surgical evaluation.

Of the appropriate cases, 55.9% (n = 306) were rated as very high priority. Not

attempting AED treatment was always rated as inappropriate for a referral. Trial

of 2 AEDs was usually rated as appropriate unless seizure-free or not fully

investigated Based on these data, a Web-based decision tool

(www.epilepsycases.com) was created.

CONCLUSIONS: Using the available evidence through 2008 and expert consensus, we

developed a Web-based decision tool that provides a guide for determining

candidacy for epilepsy surgery evaluations. The tool needs clinical validation,

and will be updated and revised regularly. This rendition of the tool is most

appropriate for those over age 12 years with focal epilepsy. The Rand/UCLA

appropriate methodology might be considered in the development of guidelines in

other areas of epilepsy care.

DOI: 10.1212/WNL.0b013e3182698c4c

PMID: 22895589 [Indexed for MEDLINE]

1304. BioData Min. 2012 Sep 7;5(1):14. doi: 10.1186/1756-0381-5-14.

Peer2ref: a peer-reviewer finding web tool that uses author disambiguation.

Andrade-Navarro MA(1), Palidwor GA, Perez-Iratxeta C.

Author information:

(1)Ottawa Hospital Research Institute, 501 Smyth Road, Ottawa, Ontario, K1H 8L6,

Canada. cpereziratxeta@gmail.com.

BACKGROUND: Reviewer and editor selection for peer review is getting harder for

authors and publishers due to the specialization onto narrower areas of research

carried by the progressive growth of the body of knowledge. Examination of the

literature facilitates finding appropriate reviewers but is time consuming and

complicated by author name ambiguities.

RESULTS: We have developed a method called peer2ref to support authors and

editors in selecting suitable reviewers for scientific manuscripts. Peer2ref

works from a text input, usually the abstract of the manuscript, from which

important concepts are extracted as keywords using a fuzzy binary relations

approach. The keywords are searched on indexed profiles of words constructed from

the bibliography attributed to authors in MEDLINE. The names of these scientists

have been previously disambiguated by coauthors identified across the whole

MEDLINE. The methods have been implemented in a web server that automatically

suggests experts for peer-review among scientists that have authored manuscripts

published during the last decade in more than 3,800 journals indexed in MEDLINE.

CONCLUSION: peer2ref web server is publicly available at

http://www.ogic.ca/projects/peer2ref/.

DOI: 10.1186/1756-0381-5-14

PMCID: PMC3499436

PMID: 22958760

1305. Bioinformatics. 2012 Sep 1;28(17):2249-55. doi: 10.1093/bioinformatics/bts426.

Epub 2012 Jul 10.

Computational prediction of N-linked glycosylation incorporating structural

properties and patterns.

Chuang GY(1), Boyington JC, Joyce MG, Zhu J, Nabel GJ, Kwong PD, Georgiev I.

Author information:

(1)Vaccine Research Center, National Institute of Allergy and Infectious

Diseases, National Institutes of Health, Bethesda, MA 20892, USA.

MOTIVATION: N-linked glycosylation occurs predominantly at the N-X-T/S motif,

where X is any amino acid except proline. Not all N-X-T/S sequons are

glycosylated, and a number of web servers for predicting N-linked glycan

occupancy using sequence and/or residue pattern information have been developed.

None of the currently available servers, however, utilizes protein structural

information for the prediction of N-glycan occupancy.

RESULTS: Here, we describe a novel classifier algorithm, NGlycPred, for the

prediction of glycan occupancy at the N-X-T/S sequons. The algorithm utilizes

both structural as well as residue pattern information and was trained on a set

of glycosylated protein structures using the Random Forest algorithm. The best

predictor achieved a balanced accuracy of 0.687 under 10-fold cross-validation on

a curated dataset of 479 N-X-T/S sequons and outperformed sequence-based

predictors when evaluated on the same dataset. The incorporation of structural

information, including local contact order, surface accessibility/composition and

secondary structure thus improves the prediction accuracy of glycan occupancy at

the N-X-T/S consensus sequon.

AVAILABILITY AND IMPLEMENTATION: NGlycPred is freely available to non-commercial

users as a web-based server at http://exon.niaid.nih.gov/nglycpred/.

DOI: 10.1093/bioinformatics/bts426

PMCID: PMC3426846

PMID: 22782545 [Indexed for MEDLINE]

1306. Gene. 2012 Sep 1;505(2):259-65. doi: 10.1016/j.gene.2012.06.014. Epub 2012 Jun

15.

TWARIT: an extremely rapid and efficient approach for phylogenetic classification

of metagenomic sequences.

Reddy RM(1), Mohammed MH, Mande SS.

Author information:

(1)Bio-sciences R&D Division, TCS Innovation Labs, Tata Research Development &

Design Centre, 54-B, Hadapsar Industrial Estate, Pune, 411013, India.

maheedhar@atc.tcs.com

Phylogenetic assignment of individual sequence reads to their respective taxa,

referred to as 'taxonomic binning', constitutes a key step of metagenomic

analysis. Existing binning methods have limitations either with respect to time

or accuracy/specificity of binning. Given these limitations, development of a

method that can bin vast amounts of metagenomic sequence data in a rapid,

efficient and computationally inexpensive manner can profoundly influence

metagenomic analysis in computational resource poor settings. We introduce

TWARIT, a hybrid binning algorithm, that employs a combination of short-read

alignment and composition-based signature sorting approaches to achieve rapid

binning rates without compromising on binning accuracy and specificity. TWARIT is

validated with simulated and real-world metagenomes and the results demonstrate

significantly lower overall binning times compared to that of existing methods.

Furthermore, the binning accuracy and specificity of TWARIT are observed to be

comparable/superior to them. A web server implementing TWARIT algorithm is

available at http://metagenomics.atc.tcs.com/Twarit/

Copyright © 2012 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.gene.2012.06.014

PMID: 22710135 [Indexed for MEDLINE]

1307. Geospat Health. 2012 Sep;6(3):S25-30.

SandflyMap: leveraging spatial data on sand fly vector distribution for disease

risk assessments.

Foley DH(1), Wilkerson RC, Dornak LL, Pecor DB, Nyari AS, Rueda LM, Long LS,

Richardson JH.

Author information:

(1)Entomology Branch, Walter Reed Army Institute of Research, 503 Robert Grant

Avenue, Silver Spring, MD 20910, USA. foleydes@si.edu

We feature SandflyMap (www.sandflymap.org), a new map service within VectorMap

(www.vectormap.org) that allows free public online access to global sand fly,

tick and mosquito collection records and habitat suitability models. Given the

short home range of sand flies, combining remote sensing and collection point

data give a powerful insight into the environmental determinants of sand fly

distribution. SandflyMap is aimed at medical entomologists, vector disease

control workers, public health officials and health planners. Data are checked

for geographical and taxonomic errors, and are comprised of vouchered specimen

information, and both published and unpublished observation data. SandflyMap uses

Microsoft Silverlight and ESRI's ArcGIS Server 10 software platform to present

disease vector data and relevant remote sensing layers in an online geographical

information system format. Users can view the locations of past vector

collections and the results of models that predict the geographic extent of

individual species. Collection records are searchable and downloadable, and Excel

collection forms with drop down lists, and Excel charts to country, are available

for data contributors to map and quality control their data. SandflyMap makes

accessible, and adds value to, the results of past sand fly collecting efforts.

We detail the workflow for entering occurrence data from the literature to

SandflyMap, using an example for sand flies from South America. We discuss the

utility of SandflyMap as a focal point to increase collaboration and to explore

the nexus between geography and vector-borne disease transmission.

DOI: 10.4081/gh.2012.119

PMID: 23032280 [Indexed for MEDLINE]

1308. J Comput Aided Mol Des. 2012 Sep;26(9):995-1003. doi: 10.1007/s10822-012-9587-5.

Epub 2012 Jul 14.

AtlasCBS: a web server to map and explore chemico-biological space.

Cortés-Cabrera A(1), Morreale A, Gago F, Abad-Zapatero C.

Author information:

(1)Departamento de Farmacología, Universidad de Alcalá, 28871, Alcalá de Henares,

Madrid, Spain.

New approaches are needed that can help decrease the unsustainable failure in

small-molecule drug discovery. Ligand Efficiency Indices (LEI) are making a great

impact on early-stage compound selection and prioritization. Given a

target-ligand database with chemical structures and associated biological

affinities/activities for a target, the AtlasCBS server generates

two-dimensional, dynamical representations of its contents in terms of LEI. These

variables allow an effective decoupling of the chemical (angular) and biological

(radial) components. BindingDB, PDBBind and ChEMBL databases are currently

implemented. Proprietary datasets can also be uploaded and compared. The utility

of this atlas-like representation in the future of drug design is highlighted

with some examples. The web server can be accessed at

http://ub.cbm.uam.es/atlascbs and https://www.ebi.ac.uk/chembl/atlascbs.

DOI: 10.1007/s10822-012-9587-5

PMID: 22798082 [Indexed for MEDLINE]

1309. Nucleic Acids Res. 2012 Sep 1;40(17):e135. Epub 2012 May 29.

Integrative analysis of gene and miRNA expression profiles with transcription

factor-miRNA feed-forward loops identifies regulators in human cancers.

Yan Z(1), Shah PK, Amin SB, Samur MK, Huang N, Wang X, Misra V, Ji H, Gabuzda D,

Li C.

Author information:

(1)Department of Biostatistics and Computational Biology, Dana-Farber Cancer

Institute and Harvard School of Public Health, Boston, MA 02215, USA.

We describe here a novel method for integrating gene and miRNA expression

profiles in cancer using feed-forward loops (FFLs) consisting of transcription

factors (TFs), miRNAs and their common target genes. The dChip-GemiNI (Gene and

miRNA Network-based Integration) method statistically ranks computationally

predicted FFLs by their explanatory power to account for differential gene and

miRNA expression between two biological conditions such as normal and cancer.

GemiNI integrates not only gene and miRNA expression data but also

computationally derived information about TF-target gene and miRNA-mRNA

interactions. Literature validation shows that the integrated modeling of

expression data and FFLs better identifies cancer-related TFs and miRNAs compared

to existing approaches. We have utilized GemiNI for analyzing six data sets of

solid cancers (liver, kidney, prostate, lung and germ cell) and found that

top-ranked FFLs account for ∼20% of transcriptome changes between normal and

cancer. We have identified common FFL regulators across multiple cancer types,

such as known FFLs consisting of MYC and miR-15/miR-17 families, and novel FFLs

consisting of ARNT, CREB1 and their miRNA partners. The results and analysis web

server are available at http://www.canevolve.org/dChip-GemiNi.

DOI: 10.1093/nar/gks395

PMCID: PMC3458521

PMID: 22645320 [Indexed for MEDLINE]

1310. Nutr Hosp. 2012 Sep-Oct;27(5):1576-82. doi: 10.3305/nh.2012.27.5.5940.

[Development of a current version of a software application for research and

practice in human nutrition (GRUNUMUR 2.0)].

[Article in Spanish]

Pérez-Llamas F(1), Garaulet M, Torralba C, Zamora S.

Author information:

(1)Departamento de Fisiología, Universidad de Murcia, Murcia, España.

frapella@um.es

The aim of this paper is the description of a new version of the software

application GRUNUMUR, a useful tool for human nutrition studies designed by the

Nutrition Research Group from the Murcia University. Similar to the first, this

second version offers the possibility to address different types of study:

dietary habits (24 h recall, 7-days dietary record and Food Frequency

Questionnaire), epidemiological, anthropometrical and clinical studies. The new

version, called GRUNUMUR 2.0, compatible with the first one, has an online help

system for all functions of the application, providing the user tasks, allows

safe storage of a virtually unlimited number of results, in an orderly and

organized way, you can retrieve it when required, through a system of backups and

scheduled maintenance and unattended (tasks performed by a server), another

advantage is its total accessibility, both from the university intranet

(www.um.es) and from the internet, it works via Web Browser

(http://senver.inf.um.es/esen), and finally, allows data to be exported to Excel

for further processing with other applications as well as publishing reports in

PDF, to deliver study participants if necessary. The new version has been

validated by comparing the extracted results with those obtained from the other

software with no significant differences for any of the variables analyzed. The

application GRUNUMUR 2.0 is a tool improved, useful and reliable for addressing

human nutrition studies.

DOI: 10.3305/nh.2012.27.5.5940

PMID: 23478708 [Indexed for MEDLINE]

1311. J Chem Inf Model. 2012 Aug 27;52(8):2310-6. doi: 10.1021/ci300245q. Epub 2012 Aug

10.

ToxAlerts: a Web server of structural alerts for toxic chemicals and compounds

with potential adverse reactions.

Sushko I(1), Salmina E, Potemkin VA, Poda G, Tetko IV.

Author information:

(1)eADMET GmbH, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany.

The article presents a Web-based platform for collecting and storing

toxicological structural alerts from literature and for virtual screening of

chemical libraries to flag potentially toxic chemicals and compounds that can

cause adverse side effects. An alert is uniquely identified by a SMARTS template,

a toxicological endpoint, and a publication where the alert was described.

Additionally, the system allows storing complementary information such as name,

comments, and mechanism of action, as well as other data. Most importantly, the

platform can be easily used for fast virtual screening of large chemical

datasets, focused libraries, or newly designed compounds against the

toxicological alerts, providing a detailed profile of the chemicals grouped by

structural alerts and endpoints. Such a facility can be used for decision making

regarding whether a compound should be tested experimentally, validated with

available QSAR models, or eliminated from consideration altogether. The

alert-based screening can also be helpful for an easier interpretation of more

complex QSAR models. The system is publicly accessible and tightly integrated

with the Online Chemical Modeling Environment (OCHEM, http://ochem.eu). The

system is open and expandable: any registered OCHEM user can introduce new

alerts, browse, edit alerts introduced by other users, and virtually screen

his/her data sets against all or selected alerts. The user sets being passed

through the structural alerts can be used at OCHEM for other typical tasks:

exporting in a wide variety of formats, development of QSAR models, additional

filtering by other criteria, etc. The database already contains almost 600

structural alerts for such endpoints as mutagenicity, carcinogenicity, skin

sensitization, compounds that undergo metabolic activation, and compounds that

form reactive metabolites and, thus, can cause adverse reactions. The ToxAlerts

platform is accessible on the Web at http://ochem.eu/alerts, and it is constantly

growing.

DOI: 10.1021/ci300245q

PMCID: PMC3640409

PMID: 22876798 [Indexed for MEDLINE]

1312. Front Plant Sci. 2012 Aug 21;3:186. doi: 10.3389/fpls.2012.00186. eCollection

2012.

Predicting and analyzing protein phosphorylation sites in plants using musite.

Yao Q(1), Gao J, Bollinger C, Thelen JJ, Xu D.

Author information:

(1)Department of Computer Science, University of Missouri Columbia, MO, USA.

Although protein phosphorylation sites can be reliably identified with

high-resolution mass spectrometry, the experimental approach is time-consuming

and resource-dependent. Furthermore, it is unlikely that an experimental approach

could catalog an entire phosphoproteome. Computational prediction of

phosphorylation sites provides an efficient and flexible way to reveal potential

phosphorylation sites and provide hypotheses in experimental design. Musite is a

tool that we previously developed to predict phosphorylation sites based solely

on protein sequence. However, it was not comprehensively applied to plants. In

this study, the phosphorylation data from Arabidopsis thaliana, B. napus, G. max,

M. truncatula, O. sativa, and Z. mays were collected for cross-species testing

and the overall plant-specific prediction as well. The results show that the

model for A. thaliana can be extended to other organisms, and the overall plant

model from Musite outperforms the current plant-specific prediction tools,

Plantphos, and PhosphAt, in prediction accuracy. Furthermore, a comparative study

of predicted phosphorylation sites across orthologs among different plants was

conducted to reveal potential evolutionary features. A bipolar distribution of

isolated, non-conserved phosphorylation sites, and highly conserved ones in terms

of the amino acid type was observed. It also shows that predicted phosphorylation

sites conserved within orthologs do not necessarily share more sequence

similarity in the flanking regions than the background, but they often inherit

protein disorder, a property that does not necessitate high sequence

conservation. Our analysis also suggests that the phosphorylation frequencies

among serine, threonine, and tyrosine correlate with their relative proportion in

disordered regions. Musite can be used as a web server (http://musite.net) or

downloaded as an open-source standalone tool (http://musite.sourceforge.net/).

DOI: 10.3389/fpls.2012.00186

PMCID: PMC3423629

PMID: 22934099

1313. Bioinformatics. 2012 Aug 15;28(16):2207-8. doi: 10.1093/bioinformatics/bts359.

Epub 2012 Jun 23.

BIDDSAT: visualizing the content of biodiversity data publishers in the Global

Biodiversity Information Facility network.

Otegui J(1), Ariño AH.

Author information:

(1)Department of Zoology and Ecology, University of Navarra, 31080 Pamplona,

Spain. jotellechea@alumni.unav.es

In any data quality workflow, data publishers must become aware of issues in

their data so these can be corrected. User feedback mechanisms provide one

avenue, while global assessments of datasets provide another. To date, there is

no publicly available tool to allow both biodiversity data institutions sharing

their data through the Global Biodiversity Information Facility network and its

potential users to assess datasets as a whole. Contributing to bridge this gap

both for publishers and users, we introduce BIoDiversity DataSets Assessment

Tool, an online tool that enables selected diagnostic visualizations on the

content of data publishers and/or their individual collections.AVAILABILITY AND

IMPLEMENTATION: The online application is accessible at

http://www.unav.es/unzyec/mzna/biddsat/ and is supported by all major browsers.

The source code is licensed under the GNU GPLv3 license

(http://www.gnu.org/licenses/gpl-3.0.txt) and is available at

https://github.com/jotegui/BIDDSAT.

DOI: 10.1093/bioinformatics/bts359

PMID: 22730433 [Indexed for MEDLINE]

1314. Bioinformatics. 2012 Aug 15;28(16):2189-90. doi: 10.1093/bioinformatics/bts343.

Epub 2012 Jun 17.

Bluues server: electrostatic properties of wild-type and mutated protein

structures.

Walsh I(1), Minervini G, Corazza A, Esposito G, Tosatto SC, Fogolari F.

Author information:

(1)Department of Biology, University of Padova, Viale G. Colombo 3, 35131 Padova,

Italy.

MOTIVATION: Electrostatic calculations are an important tool for deciphering many

functional mechanisms in proteins. Generalized Born (GB) models offer a fast and

convenient computational approximation over other implicit solvent-based

electrostatic models. Here we present a novel GB-based web server, using the

program Bluues, to calculate numerous electrostatic features including pKa-values

and surface potentials. The output is organized allowing both experts and

beginners to rapidly sift the data. A novel feature of the Bluues server is that

it explicitly allows to find electrostatic differences between wild-type and

mutant structures.

AVAILABILITY: The Bluues server, examples and extensive help files are available

for non-commercial use at URL: http://protein.bio.unipd.it/bluues/.

DOI: 10.1093/bioinformatics/bts343

PMID: 22711791 [Indexed for MEDLINE]

1315. Int J Health Geogr. 2012 Aug 14;11:33. doi: 10.1186/1476-072X-11-33.

Web-based GIS: the vector-borne disease airline importation risk (VBD-AIR) tool.

Huang Z(1), Das A, Qiu Y, Tatem AJ.

Author information:

(1)Department of Geography, University of Florida, Gainesville, FL, USA.

seenhzj@gmail.com.

BACKGROUND: Over the past century, the size and complexity of the air travel

network has increased dramatically. Nowadays, there are 29.6 million scheduled

flights per year and around 2.7 billion passengers are transported annually. The

rapid expansion of the network increasingly connects regions of endemic

vector-borne disease with the rest of the world, resulting in challenges to

health systems worldwide in terms of vector-borne pathogen importation and

disease vector invasion events. Here we describe the development of a

user-friendly Web-based GIS tool: the Vector-Borne Disease Airline Importation

Risk Tool (VBD-AIR), to help better define the roles of airports and airlines in

the transmission and spread of vector-borne diseases.

METHODS: Spatial datasets on modeled global disease and vector distributions, as

well as climatic and air network traffic data were assembled. These were combined

to derive relative risk metrics via air travel for imported infections, imported

vectors and onward transmission, and incorporated into a three-tier server

architecture in a Model-View-Controller framework with distributed GIS

components. A user-friendly web-portal was built that enables dynamic querying of

the spatial databases to provide relevant information.

RESULTS: The VBD-AIR tool constructed enables the user to explore the

interrelationships among modeled global distributions of vector-borne infectious

diseases (malaria. dengue, yellow fever and chikungunya) and international air

service routes to quantify seasonally changing risks of vector and vector-borne

disease importation and spread by air travel, forming an evidence base to help

plan mitigation strategies. The VBD-AIR tool is available at

http://www.vbd-air.com.

CONCLUSIONS: VBD-AIR supports a data flow that generates analytical results from

disparate but complementary datasets into an organized cartographical

presentation on a web map for the assessment of vector-borne disease movements on

the air travel network. The framework built provides a flexible and robust

informatics infrastructure by separating the modules of functionality through an

ontological model for vector-borne disease. The VBD‒AIR tool is designed as an

evidence base for visualizing the risks of vector-borne disease by air travel for

a wide range of users, including planners and decisions makers based in state and

local government, and in particular, those at international and domestic airports

tasked with planning for health risks and allocating limited resources.

DOI: 10.1186/1476-072X-11-33

PMCID: PMC3503742

PMID: 22892045 [Indexed for MEDLINE]

1316. BMC Res Notes. 2012 Aug 4;5:410. doi: 10.1186/1756-0500-5-410.

Distribution and prediction of catalytic domains in 2-oxoglutarate dependent

dioxygenases.

Kundu S(1).

Author information:

(1)Department of Biochemistry, Army College of Medical Sciences, Delhi Cantt.,

New Delhi 110010, India. siddharthakundu@theacms.in

BACKGROUND: The 2-oxoglutarate dependent superfamily is a diverse group of

non-haem dioxygenases, and is present in prokaryotes, eukaryotes, and archaea.

The enzymes differ in substrate preference and reaction chemistry, a factor that

precludes their classification by homology studies and electronic annotation

schemes alone. In this work, I propose and explore the rationale of using

substrates to classify structurally similar alpha-ketoglutarate dependent

enzymes.

FINDINGS: Differential catalysis in phylogenetic clades of 2-OG dependent

enzymes, is determined by the interactions of a subset of active-site amino

acids. Identifying these with existing computational methods is challenging and

not feasible for all proteins. A clustering protocol based on validated

mechanisms of catalysis of known molecules, in tandem with group specific hidden

markov model profiles is able to differentiate and sequester these enzymes.

Access to this repository is by a web server that compares user defined unknown

sequences to these pre-defined profiles and outputs a list of predicted catalytic

domains. The server is free and is accessible at the following URL

(http://comp-biol.theacms.in/H2OGpred.html).

CONCLUSIONS: The proposed stratification is a novel attempt at classifying and

predicting 2-oxoglutarate dependent function. In addition, the server will

provide researchers with a tool to compare their data to a comprehensive list of

HMM profiles of catalytic domains. This work, will aid efforts by investigators

to screen and characterize putative 2-OG dependent sequences. The profile

database will be updated at regular intervals.

DOI: 10.1186/1756-0500-5-410

PMCID: PMC3475032

PMID: 22862831 [Indexed for MEDLINE]

1317. BMC Bioinformatics. 2012 Aug 2;13:190. doi: 10.1186/1471-2105-13-190.

A web-based multi-genome synteny viewer for customized data.

Revanna KV(1), Munro D, Gao A, Chiu CC, Pathak A, Dong Q.

Author information:

(1)Department of Biological Sciences, University of North Texas, Denton, TX

76203, USA.

BACKGROUND: Web-based synteny visualization tools are important for sharing data

and revealing patterns of complicated genome conservation and rearrangements.

Such tools should allow biologists to upload genomic data for their own analysis.

This requirement is critical because individual biologists are generating large

amounts of genomic sequences that quickly overwhelm any centralized web resources

to collect and display all those data. Recently, we published a web-based synteny

viewer, GSV, which was designed to satisfy the above requirement. However, GSV

can only compare two genomes at a given time. Extending the functionality of GSV

to visualize multiple genomes is important to meet the increasing demand of the

research community.

RESULTS: We have developed a multi-Genome Synteny Viewer (mGSV). Similar to GSV,

mGSV is a web-based tool that allows users to upload their own genomic data files

for visualization. Multiple genomes can be presented in a single integrated view

with an enhanced user interface. Users can navigate through all the selected

genomes in either pairwise or multiple viewing mode to examine conserved genomic

regions as well as the accompanying genome annotations. Besides serving users who

manually interact with the web server, mGSV also provides Web Services for

machine-to-machine communication to accept data sent by other remote resources.

The entire mGSV package can also be downloaded for easy local installation.

CONCLUSIONS: mGSV significantly enhances the original functionalities of GSV. A

web server hosting mGSV is provided at http://cas-bioinfo.cas.unt.edu/mgsv.

DOI: 10.1186/1471-2105-13-190

PMCID: PMC3430549

PMID: 22856879 [Indexed for MEDLINE]

1318. BMC Bioinformatics. 2012 Aug 2;13:188. doi: 10.1186/1471-2105-13-188.

A resource for benchmarking the usefulness of protein structure models.

Carbajo D(1), Tramontano A.

Author information:

(1)Department of Physics, Sapienza University of Rome, P,le A, Moro, 5, 00185

Rome, Italy.

BACKGROUND: Increasingly, biologists and biochemists use computational tools to

design experiments to probe the function of proteins and/or to engineer them for

a variety of different purposes. The most effective strategies rely on the

knowledge of the three-dimensional structure of the protein of interest. However

it is often the case that an experimental structure is not available and that

models of different quality are used instead. On the other hand, the relationship

between the quality of a model and its appropriate use is not easy to derive in

general, and so far it has been analyzed in detail only for specific application.

RESULTS: This paper describes a database and related software tools that allow

testing of a given structure based method on models of a protein representing

different levels of accuracy. The comparison of the results of a computational

experiment on the experimental structure and on a set of its decoy models will

allow developers and users to assess which is the specific threshold of accuracy

required to perform the task effectively.

CONCLUSIONS: The ModelDB server automatically builds decoy models of different

accuracy for a given protein of known structure and provides a set of useful

tools for their analysis. Pre-computed data for a non-redundant set of deposited

protein structures are available for analysis and download in the ModelDB

database. IMPLEMENTATION, AVAILABILITY AND REQUIREMENTS: Project name: A resource

for benchmarking the usefulness of protein structure models. Project home page:

http://bl210.caspur.it/MODEL-DB/MODEL-DB\_web/MODindex.php.Operating system(s):

Platform independent. Programming language: Perl-BioPerl (program); mySQL, Perl

DBI and DBD modules (database); php, JavaScript, Jmol scripting (web server).

Other requirements: Java Runtime Environment v1.4 or later, Perl, BioPerl, CPAN

modules, HHsearch, Modeller, LGA, NCBI Blast package, DSSP, Speedfill (Surfnet)

and PSAIA. License: Free. Any restrictions to use by non-academics: No.

DOI: 10.1186/1471-2105-13-188

PMCID: PMC3473236

PMID: 22856649 [Indexed for MEDLINE]

1319. Am J Physiol Renal Physiol. 2012 Aug 1;303(3):F366-72. doi:

10.1152/ajprenal.00147.2012. Epub 2012 May 30.

An online tool for calculation of free-energy balance for the renal inner

medulla.

Vilbig RL(1), Sarkar A, Zischkau J, Knepper MA, Pisitkun T.

Author information:

(1)Epithelial Systems Biology Laboratory, National Heart, Lung and Blood

Institute, National Institutes of Health, Bethesda, MD 20892-1603, USA.

Concentrating models of the renal inner medulla can be classified according to

external free-energy balance into passive models (positive values) and models

that require an external energy source (negative values). Here we introduce an

online computational tool that implements the equations of Stephenson and

colleagues (Stephenson JL, Tewarson RP, Mejia R. Proc Natl Acad Sci USA 71:

1618-1622, 1974) to calculate external free-energy balance at steady state for

the inner medulla (http://helixweb.nih.gov/ESBL/FreeEnergy). Here "external

free-energy balance" means the sum of free-energy flows in all streams entering

and leaving the inner medulla. The program first assures steady-state mass

balance for all components and then tallies net external free-energy balance for

the selected flow conditions. Its use is illustrated by calculating external

free-energy balance for an example of the passive concentrating model taken from

the original paper by Kokko and Rector (Kokko JP, Rector FC Jr. Kidney Int 2:

214-223, 1972).

DOI: 10.1152/ajprenal.00147.2012

PMCID: PMC3433863

PMID: 22647629 [Indexed for MEDLINE]

1320. Bioinformatics. 2012 Aug 1;28(15):2067-8. doi: 10.1093/bioinformatics/bts325.

Epub 2012 Jun 4.

SYNCSA--R tool for analysis of metacommunities based on functional traits and

phylogeny of the community components.

Debastiani VJ(1), Pillar VD.

Author information:

(1)Programa de Pós-Graduação em Ecologia, Universidade Federal do Rio Grande do

Sul, Porto Alegre, RS 91501-970, Brazil. vanderleidebastiani@yahoo.com.br

SUMMARY: SYNCSA is an R package for the analysis of metacommunities based on

functional traits and phylogeny of the community components. It offers tools to

calculate several matrix correlations that express trait-convergence assembly

patterns, trait-divergence assembly patterns and phylogenetic signal in

functional traits at the species pool level and at the metacommunity level.

AVAILABILITY AND IMPLEMENTATION: SYNCSA is a package for the R environment, under

a GPL-2 open-source license and freely available on CRAN official web server for

R (http://cran.r-project.org).

CONTACT: vanderleidebastiani@yahoo.com.br.

DOI: 10.1093/bioinformatics/bts325

PMID: 22668789 [Indexed for MEDLINE]

1321. Bioinformatics. 2012 Aug 1;28(15):2078-9. doi: 10.1093/bioinformatics/bts321.

Epub 2012 Jun 1.

POOL server: machine learning application for functional site prediction in

proteins.

Somarowthu S(1), Ondrechen MJ.

Author information:

(1)Department of Chemistry and Chemical Biology, Northeastern University, Boston,

MA 02115, USA.

SUMMARY: We present an automated web server for partial order optimum likelihood

(POOL), a machine learning application that combines computed electrostatic and

geometric information for high-performance prediction of catalytic residues from

3D structures. Input features consist of THEMATICS electrostatics data and pocket

information from ConCavity. THEMATICS measures deviation from typical, sigmoidal

titration behavior to identify functionally important residues and ConCavity

identifies binding pockets by analyzing the surface geometry of protein

structures. Both THEMATICS and ConCavity (structure only) do not require the

query protein to have any sequence or structure similarity to other proteins.

Hence, POOL is applicable to proteins with novel folds and engineered proteins.

As an additional option for cases where sequence homologues are available, users

can include evolutionary information from INTREPID for enhanced accuracy in site

prediction.

AVAILABILITY: The web site is free and open to all users with no login

requirements at http://www.pool.neu.edu.

CONTACT: m.ondrechen@neu.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/bts321

PMCID: PMC3400966

PMID: 22661648 [Indexed for MEDLINE]

1322. Bioinformatics. 2012 Aug 1;28(15):2076-7. doi: 10.1093/bioinformatics/bts320.

Epub 2012 May 29.

AS-EAST: a functional annotation tool for putative proteins encoded by

alternatively spliced transcripts.

Shionyu M(1), Takahashi K, Go M.

Author information:

(1)Department of Computer Bioscience, Nagahama Institute of Bio-Science and

Technology, 1266, Tamura-cho, Nagahama, Shiga 526-0829, Japan.

m\_shionyu@nagahama-i-bio.ac.jp

SUMMARY: Alternative Splicing Effects ASsessment Tools (AS-EAST) is an online

tool for the functional annotation of putative proteins encoded by transcripts

generated by alternative splicing (AS). When provided with a transcript sequence,

AS-EAST identifies regions altered by AS events in the putative protein sequence

encoded by the transcript. Users can evaluate the predicted function of the

putative protein by inspecting whether functional domains are included in the

altered regions. Moreover, users can infer the loss of inter-molecular

interactions in the protein network according to whether the AS events affect

interaction residues observed in the 3D structure of the reference isoform. The

information obtained from AS-EAST will help to design experimental analyses for

the functional significance of novel splice isoforms.

AVAILABILITY: The online tool is freely available at

http://as-alps.nagahama-i-bio.ac.jp/ASEAST/.

CONTACT: m\_shionyu@nagahama-i-bio.ac.jp.

DOI: 10.1093/bioinformatics/bts320

PMCID: PMC3400965

PMID: 22645168 [Indexed for MEDLINE]

1323. Bioinformatics. 2012 Aug 1;28(15):2074-5. doi: 10.1093/bioinformatics/bts310.

Epub 2012 May 23.

DoGSiteScorer: a web server for automatic binding site prediction, analysis and

druggability assessment.

Volkamer A(1), Kuhn D, Rippmann F, Rarey M.

Author information:

(1)Center for Bioinformatics, University of Hamburg, Bundesstr, Germany.

MOTIVATION: Many drug discovery projects fail because the underlying target is

finally found to be undruggable. Progress in structure elucidation of proteins

now opens up a route to automatic structure-based target assessment.

DoGSiteScorer is a newly developed automatic tool combining pocket prediction,

characterization and druggability estimation and is now available through a web

server.

AVAILABILITY: The DoGSiteScorer web server is freely available for academic use

at http://dogsite.zbh.uni-hamburg.de

CONTACT: rarey@zbh.uni-hamburg.de.

DOI: 10.1093/bioinformatics/bts310

PMID: 22628523 [Indexed for MEDLINE]

1324. Bioinformatics. 2012 Aug 1;28(15):2072-3. doi: 10.1093/bioinformatics/bts302.

Epub 2012 May 21.

SALIGN: a web server for alignment of multiple protein sequences and structures.

Braberg H(1), Webb BM, Tjioe E, Pieper U, Sali A, Madhusudhan MS.

Author information:

(1)Department of Cellular and Molecular Pharmacology, University of California,

San Francisco, CA 94158, USA.

SUMMARY: Accurate alignment of protein sequences and/or structures is crucial for

many biological analyses, including functional annotation of proteins,

classifying protein sequences into families, and comparative protein structure

modeling. Described here is a web interface to SALIGN, the versatile protein

multiple sequence/structure alignment module of MODELLER. The web server

automatically determines the best alignment procedure based on the inputs, while

allowing the user to override default parameter values. Multiple alignments are

guided by a dendrogram computed from a matrix of all pairwise alignment scores.

When aligning sequences to structures, SALIGN uses structural environment

information to place gaps optimally. If two multiple sequence alignments of

related proteins are input to the server, a profile-profile alignment is

performed. All features of the server have been previously optimized for

accuracy, especially in the contexts of comparative modeling and identification

of interacting protein partners.

AVAILABILITY: The SALIGN web server is freely accessible to the academic

community at http://salilab.org/salign. SALIGN is a module of the MODELLER

software, also freely available to academic users (http://salilab.org/modeller).

CONTACT: sali@salilab.org; madhusudhan@bii.a-star.edu.sg.

DOI: 10.1093/bioinformatics/bts302

PMCID: PMC3400954

PMID: 22618536 [Indexed for MEDLINE]

1325. Proteins. 2012 Aug;80(8):2080-8. doi: 10.1002/prot.24100. Epub 2012 May 25.

A new size-independent score for pairwise protein structure alignment and its

application to structure classification and nucleic-acid binding prediction.

Yang Y(1), Zhan J, Zhao H, Zhou Y.

Author information:

(1)Indiana University School of Informatics, Indiana University-Purdue

University, Indianapolis, Indiana 46202, USA.

A structure alignment program aligns two structures by optimizing a scoring

function that measures structural similarity. It is highly desirable that such

scoring function is independent of the sizes of proteins in comparison so that

the significance of alignment across different sizes of the protein regions

aligned is comparable. Here, we developed a new score called SP-score that fixes

the cutoff distance at 4 Å and removed the size dependence using a normalization

prefactor. We further built a program called SPalign that optimizes SP-score for

structure alignment. SPalign was applied to recognize proteins within the same

structure fold and having the same function of DNA or RNA binding. For fold

discrimination, SPalign improves sensitivity over TMalign for the chain-level

comparison by 12% and over DALI for the domain-level comparison by 13% at the

same specificity of 99.6%. The difference between TMalign and SPalign at the

chain level is due to the inability of TMalign to detect single domain similarity

between multidomain proteins. For recognizing nucleic acid binding proteins,

SPalign consistently improves over TMalign by 12% and DALI by 31% in average

value of Mathews correlation coefficients for four datasets. SPalign with default

setting is 14% faster than TMalign. SPalign is expected to be useful for function

prediction and comparing structures with or without domains defined. The source

code for SPalign and the server are available at

http://sparks.informatics.iupui.edu.

Copyright © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24100

PMCID: PMC3393833

PMID: 22522696 [Indexed for MEDLINE]

1326. Source Code Biol Med. 2012 Jul 30;7(1):8. doi: 10.1186/1751-0473-7-8.

JobCenter: an open source, cross-platform, and distributed job queue management

system optimized for scalability and versatility.

Jaschob D(1), Riffle M.

Author information:

(1)Department of Biochemistry, University of Washington, 1705 NE Pacific St, Box

357350, Seattle, WA, 98195, USA. mriffle@uw.edu.

BACKGROUND: Laboratories engaged in computational biology or bioinformatics

frequently need to run lengthy, multistep, and user-driven computational jobs.

Each job can tie up a computer for a few minutes to several days, and many

laboratories lack the expertise or resources to build and maintain a dedicated

computer cluster.

RESULTS: JobCenter is a client-server application and framework for job

management and distributed job execution. The client and server components are

both written in Java and are cross-platform and relatively easy to install. All

communication with the server is client-driven, which allows worker nodes to run

anywhere (even behind external firewalls or "in the cloud") and provides inherent

load balancing. Adding a worker node to the worker pool is as simple as dropping

the JobCenter client files onto any computer and performing basic configuration,

which provides tremendous ease-of-use, flexibility, and limitless horizontal

scalability. Each worker installation may be independently configured, including

the types of jobs it is able to run. Executed jobs may be written in any language

and may include multistep workflows.

CONCLUSIONS: JobCenter is a versatile and scalable distributed job management

system that allows laboratories to very efficiently distribute all computational

work among available resources. JobCenter is freely available at

http://code.google.com/p/jobcenter/.

DOI: 10.1186/1751-0473-7-8

PMCID: PMC3494518

PMID: 22846423

1327. BMC Syst Biol. 2012 Jul 26;6:91. doi: 10.1186/1752-0509-6-91.

Condor-COPASI: high-throughput computing for biochemical networks.

Kent E(1), Hoops S, Mendes P.

Author information:

(1)Doctoral Training Centre in Integrative Systems Biology, Manchester Institute

of Biotechnology, The University of Manchester, 131 Princess Street, Manchester

M1 7DN, UK.

BACKGROUND: Mathematical modelling has become a standard technique to improve our

understanding of complex biological systems. As models become larger and more

complex, simulations and analyses require increasing amounts of computational

power. Clusters of computers in a high-throughput computing environment can help

to provide the resources required for computationally expensive model analysis.

However, exploiting such a system can be difficult for users without the

necessary expertise.

RESULTS: We present Condor-COPASI, a server-based software tool that integrates

COPASI, a biological pathway simulation tool, with Condor, a high-throughput

computing environment. Condor-COPASI provides a web-based interface, which makes

it extremely easy for a user to run a number of model simulation and analysis

tasks in parallel. Tasks are transparently split into smaller parts, and

submitted for execution on a Condor pool. Result output is presented to the user

in a number of formats, including tables and interactive graphical displays.

CONCLUSIONS: Condor-COPASI can effectively use a Condor high-throughput computing

environment to provide significant gains in performance for a number of model

simulation and analysis tasks. Condor-COPASI is free, open source software,

released under the Artistic License 2.0, and is suitable for use by any

institution with access to a Condor pool. Source code is freely available for

download at http://code.google.com/p/condor-copasi/, along with full instructions

on deployment and usage.

DOI: 10.1186/1752-0509-6-91

PMCID: PMC3527284

PMID: 22834945 [Indexed for MEDLINE]

1328. Nat Protoc. 2012 Jul 26;7(8):1551-68. doi: 10.1038/nprot.2012.088.

A complete workflow for the analysis of full-size ChIP-seq (and similar) data

sets using peak-motifs.

Thomas-Chollier M(1), Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J.

Author information:

(1)Department of Computational Molecular Biology, Max Planck Institute for

Molecular Genetics, Berlin, Germany. thomas-c@molgen.mpg.de

This protocol explains how to use the online integrated pipeline 'peak-motifs'

(http://rsat.ulb.ac.be/rsat/) to predict motifs and binding sites in full-size

peak sets obtained by chromatin immunoprecipitation-sequencing (ChIP-seq) or

related technologies. The workflow combines four time- and memory-efficient motif

discovery algorithms to extract significant motifs from the sequences. Discovered

motifs are compared with databases of known motifs to identify potentially bound

transcription factors. Sequences are scanned to predict transcription factor

binding sites and analyze their enrichment and positional distribution relative

to peak centers. Peaks and binding sites are exported as BED tracks that can be

uploaded into the University of California Santa Cruz (UCSC) genome browser for

visualization in the genomic context. This protocol is illustrated with the

analysis of a set of 6,000 peaks (8 Mb in total) bound by the Drosophila

transcription factor Krüppel. The complete workflow is achieved in about 25 min

of computational time on the Regulatory Sequence Analysis Tools (RSAT) Web

server. This protocol can be followed in about 1 h.

DOI: 10.1038/nprot.2012.088

PMID: 22836136 [Indexed for MEDLINE]

1329. BMC Syst Biol. 2012 Jul 23;6:90. doi: 10.1186/1752-0509-6-90.

Multiple independent analyses reveal only transcription factors as an enriched

functional class associated with microRNAs.

Croft L(1), Szklarczyk D, Jensen LJ, Gorodkin J.

Author information:

(1)Center for Non-coding RNA in Technology and Health, Division of Genetics and

Bioinformatics, IBHV, University of Copenhagen, Copenhagen, Denmark.

BACKGROUND: Transcription factors (TFs) have long been known to be principally

activators of transcription in eukaryotes and prokaryotes. The growing awareness

of the ubiquity of microRNAs (miRNAs) as suppressive regulators in eukaryotes,

suggests the possibility of a mutual, preferential, self-regulatory connectivity

between miRNAs and TFs. Here we investigate the connectivity from TFs and miRNAs

to other genes and each other using text mining, TF promoter binding site and 6

different miRNA binding site prediction methods.

RESULTS: In the first approach text mining of PubMed abstracts reveal

statistically significant associations between miRNAs and both TFs and signal

transduction gene classes. Secondly, prediction of miRNA targets in human and

mouse 3'UTRs show enrichment only for TFs but not consistently across prediction

methods for signal transduction or other gene classes. Furthermore, a random

sample of 986 TarBase entries was scored for experimental evidence by manual

inspection of the original papers, and enrichment for TFs was observed to

increase with score. Low-scoring TarBase entries, where experimental evidence is

anticorrelated miRNA:mRNA expression with predicted miRNA targets, appear not to

select for real miRNA targets to any degree. Our manually validated text-mining

results also suggests that miRNAs may be activated by more TFs than other classes

of genes, as 7% of miRNA:TF co-occurrences in the literature were TFs activating

miRNAs. This was confirmed when thirdly, we found enrichment for predicted,

conserved TF binding sites in miRNA and TF genes compared to other gene classes.

CONCLUSIONS: We see enrichment of connections between miRNAs and TFs using

several independent methods, suggestive of a network of mutual activating and

suppressive regulation. We have also built regulatory networks (containing 2- and

3-loop motifs) for mouse and human using predicted miRNA and TF binding sites and

we have developed a web server to search and display these loops, available for

the community at http://rth.dk/resources/tfmirloop.

DOI: 10.1186/1752-0509-6-90

PMCID: PMC3430561

PMID: 22824421 [Indexed for MEDLINE]

1330. Nat Protoc. 2012 Jul 19;7(8):1511-22. doi: 10.1038/nprot.2012.085.

Template-based protein structure modeling using the RaptorX web server.

Källberg M(1), Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J.

Author information:

(1)Toyota Technological Institute at Chicago, Chicago, Illinois, USA.

A key challenge of modern biology is to uncover the functional role of the

protein entities that compose cellular proteomes. To this end, the availability

of reliable three-dimensional atomic models of proteins is often crucial. This

protocol presents a community-wide web-based method using RaptorX

(http://raptorx.uchicago.edu/) for protein secondary structure prediction,

template-based tertiary structure modeling, alignment quality assessment and

sophisticated probabilistic alignment sampling. RaptorX distinguishes itself from

other servers by the quality of the alignment between a target sequence and one

or multiple distantly related template proteins (especially those with sparse

sequence profiles) and by a novel nonlinear scoring function and a

probabilistic-consistency algorithm. Consequently, RaptorX delivers high-quality

structural models for many targets with only remote templates. At present, it

takes RaptorX ~35 min to finish processing a sequence of 200 amino acids. Since

its official release in August 2011, RaptorX has processed ~6,000 sequences

submitted by ~1,600 users from around the world.

DOI: 10.1038/nprot.2012.085

PMCID: PMC4730388

PMID: 22814390 [Indexed for MEDLINE]

1331. BioData Min. 2012 Jul 16;5(1):7. doi: 10.1186/1756-0381-5-7.

'MicroRNA Targets', a new AthaMap web-tool for genome-wide identification of

miRNA targets in Arabidopsis thaliana.

Bülow L(1), Bolívar JC, Ruhe J, Brill Y, Hehl R.

Author information:

(1)Institut für Genetik, Technische Universität Braunschweig, Spielmannstr,

38106, Braunschweig, Germany. r.hehl@tu-braunschweig.de.

BACKGROUND: The AthaMap database generates a genome-wide map for putative

transcription factor binding sites for A. thaliana. When analyzing

transcriptional regulation using AthaMap it may be important to learn which genes

are also post-transcriptionally regulated by inhibitory RNAs. Therefore, a

unified database for transcriptional and post-transcriptional regulation will be

highly useful for the analysis of gene expression regulation.

METHODS: To identify putative microRNA target sites in the genome of A. thaliana,

processed mature miRNAs from 243 annotated miRNA genes were used for screening

with the psRNATarget web server. Positional information, target genes and the

psRNATarget score for each target site were annotated to the AthaMap database.

Furthermore, putative target sites for small RNAs from seven small RNA

transcriptome datasets were used to determine small RNA target sites within the

A. thaliana genome.

RESULTS: Putative 41,965 genome wide miRNA target sites and 10,442 miRNA target

genes were identified in the A. thaliana genome. Taken together with genes

targeted by small RNAs from small RNA transcriptome datasets, a total of 16,600

A. thaliana genes are putatively regulated by inhibitory RNAs. A novel web-tool,

'MicroRNA Targets', was integrated into AthaMap which permits the identification

of genes predicted to be regulated by selected miRNAs. The predicted target genes

are displayed with positional information and the psRNATarget score of the target

site. Furthermore, putative target sites of small RNAs from selected tissue

datasets can be identified with the new 'Small RNA Targets' web-tool.

CONCLUSIONS: The integration of predicted miRNA and small RNA target sites with

transcription factor binding sites will be useful for AthaMap-assisted gene

expression analysis. URL: http://www.athamap.de/

DOI: 10.1186/1756-0381-5-7

PMCID: PMC3410767

PMID: 22800758

1332. Bioinformatics. 2012 Jul 15;28(14):1925-7. doi: 10.1093/bioinformatics/bts282.

Epub 2012 May 9.

CPSS: a computational platform for the analysis of small RNA deep sequencing

data.

Zhang Y(1), Xu B, Yang Y, Ban R, Zhang H, Jiang X, Cooke HJ, Xue Y, Shi Q.

Author information:

(1)Hefei National Laboratory for Physical Sciences at Microscale and School of

Life Sciences, University of Science and Technology of China, Hefei 230027,

China.

Next generation sequencing (NGS) techniques have been widely used to document the

small ribonucleic acids (RNAs) implicated in a variety of biological,

physiological and pathological processes. An integrated computational tool is

needed for handling and analysing the enormous datasets from small RNA deep

sequencing approach. Herein, we present a novel web server, CPSS (a computational

platform for the analysis of small RNA deep sequencing data), designed to

completely annotate and functionally analyse microRNAs (miRNAs) from NGS data on

one platform with a single data submission. Small RNA NGS data can be submitted

to this server with analysis results being returned in two parts: (i) annotation

analysis, which provides the most comprehensive analysis for small RNA

transcriptome, including length distribution and genome mapping of sequencing

reads, small RNA quantification, prediction of novel miRNAs, identification of

differentially expressed miRNAs, piwi-interacting RNAs and other non-coding small

RNAs between paired samples and detection of miRNA editing and modifications and

(ii) functional analysis, including prediction of miRNA targeted genes by

multiple tools, enrichment of gene ontology terms, signalling pathway involvement

and protein-protein interaction analysis for the predicted genes. CPSS, a

ready-to-use web server that integrates most functions of currently available

bioinformatics tools, provides all the information wanted by the majority of

users from small RNA deep sequencing datasets.AVAILABILITY: CPSS is implemented

in PHP/PERL+MySQL+R and can be freely accessed at

http://mcg.ustc.edu.cn/db/cpss/index.html or

http://mcg.ustc.edu.cn/sdap1/cpss/index.html.

DOI: 10.1093/bioinformatics/bts282

PMID: 22576177 [Indexed for MEDLINE]

1333. BMC Res Notes. 2012 Jul 10;5:351. doi: 10.1186/1756-0500-5-351.

ngLOC: software and web server for predicting protein subcellular localization in

prokaryotes and eukaryotes.

King BR(1), Vural S, Pandey S, Barteau A, Guda C.

Author information:

(1)Department of Computer Science, Bucknell University, One Dent Drive,

Lewisburg, PA 17837, USA.

BACKGROUND: Understanding protein subcellular localization is a necessary

component toward understanding the overall function of a protein. Numerous

computational methods have been published over the past decade, with varying

degrees of success. Despite the large number of published methods in this area,

only a small fraction of them are available for researchers to use in their own

studies. Of those that are available, many are limited by predicting only a small

number of organelles in the cell. Additionally, the majority of methods predict

only a single location for a sequence, even though it is known that a large

fraction of the proteins in eukaryotic species shuttle between locations to carry

out their function.

FINDINGS: We present a software package and a web server for predicting the

subcellular localization of protein sequences based on the ngLOC method. ngLOC is

an n-gram-based Bayesian classifier that predicts subcellular localization of

proteins both in prokaryotes and eukaryotes. The overall prediction accuracy

varies from 89.8% to 91.4% across species. This program can predict 11 distinct

locations each in plant and animal species. ngLOC also predicts 4 and 5 distinct

locations on gram-positive and gram-negative bacterial datasets, respectively.

CONCLUSIONS: ngLOC is a generic method that can be trained by data from a variety

of species or classes for predicting protein subcellular localization. The

standalone software is freely available for academic use under GNU GPL, and the

ngLOC web server is also accessible at http://ngloc.unmc.edu.

DOI: 10.1186/1756-0500-5-351

PMCID: PMC3532370

PMID: 22780965 [Indexed for MEDLINE]

1334. BMC Bioinformatics. 2012 Jul 3;13:157. doi: 10.1186/1471-2105-13-157.

Minimalist ensemble algorithms for genome-wide protein localization prediction.

Lin JR(1), Mondal AM, Liu R, Hu J.

Author information:

(1)Department of Computer Science and Engineering, University of South Carolina,

Columbia, SC 29208, USA.

BACKGROUND: Computational prediction of protein subcellular localization can

greatly help to elucidate its functions. Despite the existence of dozens of

protein localization prediction algorithms, the prediction accuracy and coverage

are still low. Several ensemble algorithms have been proposed to improve the

prediction performance, which usually include as many as 10 or more individual

localization algorithms. However, their performance is still limited by the

running complexity and redundancy among individual prediction algorithms.

RESULTS: This paper proposed a novel method for rational design of minimalist

ensemble algorithms for practical genome-wide protein subcellular localization

prediction. The algorithm is based on combining a feature selection based filter

and a logistic regression classifier. Using a novel concept of contribution

scores, we analyzed issues of algorithm redundancy, consensus mistakes, and

algorithm complementarity in designing ensemble algorithms. We applied the

proposed minimalist logistic regression (LR) ensemble algorithm to two

genome-wide datasets of Yeast and Human and compared its performance with current

ensemble algorithms. Experimental results showed that the minimalist ensemble

algorithm can achieve high prediction accuracy with only 1/3 to 1/2 of individual

predictors of current ensemble algorithms, which greatly reduces computational

complexity and running time. It was found that the high performance ensemble

algorithms are usually composed of the predictors that together cover most of

available features. Compared to the best individual predictor, our ensemble

algorithm improved the prediction accuracy from AUC score of 0.558 to 0.707 for

the Yeast dataset and from 0.628 to 0.646 for the Human dataset. Compared with

popular weighted voting based ensemble algorithms, our classifier-based ensemble

algorithms achieved much better performance without suffering from inclusion of

too many individual predictors.

CONCLUSIONS: We proposed a method for rational design of minimalist ensemble

algorithms using feature selection and classifiers. The proposed minimalist

ensemble algorithm based on logistic regression can achieve equal or better

prediction performance while using only half or one-third of individual

predictors compared to other ensemble algorithms. The results also suggested that

meta-predictors that take advantage of a variety of features by combining

individual predictors tend to achieve the best performance. The LR ensemble

server and related benchmark datasets are available at

http://mleg.cse.sc.edu/LRensemble/cgi-bin/predict.cgi.

DOI: 10.1186/1471-2105-13-157

PMCID: PMC3426488

PMID: 22759391 [Indexed for MEDLINE]

1335. Amino Acids. 2012 Jul;43(1):447-55. doi: 10.1007/s00726-011-1100-2. Epub 2011 Oct

7.

Predicting protein sumoylation sites from sequence features.

Teng S(1), Luo H, Wang L.

Author information:

(1)Department of Genetics and Biochemistry, Clemson University, Clemson, SC

29634, USA.

Protein sumoylation is a post-translational modification that plays an important

role in a wide range of cellular processes. Small ubiquitin-related modifier

(SUMO) can be covalently and reversibly conjugated to the sumoylation sites of

target proteins, many of which are implicated in various human genetic disorders.

The accurate prediction of protein sumoylation sites may help biomedical

researchers to design their experiments and understand the molecular mechanism of

protein sumoylation. In this study, a new machine learning approach has been

developed for predicting sumoylation sites from protein sequence information.

Random forests (RFs) and support vector machines (SVMs) were trained with the

data collected from the literature. Domain-specific knowledge in terms of

relevant biological features was used for input vector encoding. It was shown

that RF classifier performance was affected by the sequence context of

sumoylation sites, and 20 residues with the core motif ΨKXE in the middle

appeared to provide enough context information for sumoylation site prediction.

The RF classifiers were also found to outperform SVM models for predicting

protein sumoylation sites from sequence features. The results suggest that the

machine learning approach gives rise to more accurate prediction of protein

sumoylation sites than the other existing methods. The accurate classifiers have

been used to develop a new web server, called seeSUMO

(http://bioinfo.ggc.org/seesumo/), for sequence-based prediction of protein

sumoylation sites.

DOI: 10.1007/s00726-011-1100-2

PMID: 21986959 [Indexed for MEDLINE]

1336. Ann Occup Hyg. 2012 Jul;56(5):525-41. doi: 10.1093/annhyg/mer113. Epub 2012 Jan

20.

Stoffenmanager Nano version 1.0: a web-based tool for risk prioritization of

airborne manufactured nano objects.

Van Duuren-Stuurman B(1), Vink SR, Verbist KJ, Heussen HG, Brouwer DH, Kroese DE,

Van Niftrik MF, Tielemans E, Fransman W.

Author information:

(1)TNO Innovation for Life, Utrechtseweg 48, PO Box 350, 3700 AJ Zeist, The

Netherlands. birgit.vanduuren@tno.nl

Stoffenmanager Nano (version 1.0) is a risk-banding tool developed for employers

and employees to prioritize health risks occurring as a result of exposure to

manufactured nano objects (MNOs) for a broad range of worker scenarios and to

assist implementation of control measures to reduce exposure levels. In order to

prioritize the health risks, the Stoffenmanager Nano combines the available

hazard information of a substance with a qualitative estimate of potential for

inhalation exposure. The development of the Stoffenmanager Nano started with a

review of the available literature on control banding. Input parameters for the

hazard assessment of MNOs were selected based on the availability of these

parameters in, for instance, Safety Data Sheets or product information sheets.

The conceptual exposure model described by Schneider et al. (2011) was used as

the starting point for exposure banding. During the development of the

Stoffenmanager Nano tool, the precautionary principle was applied to deal with

the uncertainty regarding hazard and exposure assessment of MNOs. Subsequently,

the model was converted into an online tool (http://nano.stoffenmanager.nl),

tested, and reviewed by a number of companies. In this paper, we describe the

Stoffenmanager Nano. This tool offers a practical approach for risk

prioritization in exposure situations where quantitative risk assessment is

currently not possible. Updates of this first version are anticipated as more

data become available in the future.

DOI: 10.1093/annhyg/mer113

PMID: 22267129 [Indexed for MEDLINE]

1337. Bioinformatics. 2012 Jul 1;28(13):1805-6. doi: 10.1093/bioinformatics/bts251.

Epub 2012 Apr 27.

DAVID-WS: a stateful web service to facilitate gene/protein list analysis.

Jiao X(1), Sherman BT, Huang da W, Stephens R, Baseler MW, Lane HC, Lempicki RA.

Author information:

(1)Laboratory of Immunopathogenesis and Bioinformatics, National Cancer Institute

at Frederick, Frederick, MD 21702, USA. xiaoli.jiao@nih.gov

SUMMARY: The database for annotation, visualization and integrated discovery

(DAVID), which can be freely accessed at http://david.abcc.ncifcrf.gov/, is a

web-based online bioinformatics resource that aims to provide tools for the

functional interpretation of large lists of genes/proteins. It has been used by

researchers from more than 5000 institutes worldwide, with a daily submission

rate of ∼1200 gene lists from ∼400 unique researchers, and has been cited by more

than 6000 scientific publications. However, the current web interface does not

support programmatic access to DAVID, and the uniform resource locator

(URL)-based application programming interface (API) has a limit on URL size and

is stateless in nature as it uses URL request and response messages to

communicate with the server, without keeping any state-related details. DAVID-WS

(web service) has been developed to automate user tasks by providing stateful web

services to access DAVID programmatically without the need for human

interactions.

AVAILABILITY: The web service and sample clients (written in Java, Perl, Python

and Matlab) are made freely available under the DAVID License at

http://david.abcc.ncifcrf.gov/content.jsp?file=WS.html.

DOI: 10.1093/bioinformatics/bts251

PMCID: PMC3381967

PMID: 22543366 [Indexed for MEDLINE]

1338. Genomics. 2012 Jul;100(1):8-13. doi: 10.1016/j.ygeno.2012.05.007. Epub 2012 May

15.

EXP-PAC: providing comparative analysis and storage of next generation gene

expression data.

Church PC(1), Goscinski A, Lefèvre C.

Author information:

(1)School of Information Technology, Faculty of Science and Technology, Deakin

University, Geelong VIC 3127, Australia. pcc@deakin.edu.au

Microarrays and more recently RNA sequencing has led to an increase in available

gene expression data. How to manage and store this data is becoming a key issue.

In response we have developed EXP-PAC, a web based software package for storage,

management and analysis of gene expression and sequence data. Unique to this

package is SQL based querying of gene expression data sets, distributed

normalization of raw gene expression data and analysis of gene expression data

across experiments and species. This package has been populated with lactation

data in the international milk genomic consortium web portal

(http://milkgenomics.org/). Source code is also available which can be hosted on

a Windows, Linux or Mac APACHE server connected to a private or public network

(http://mamsap.it.deakin.edu.au/~pcc/Release/EXP\_PAC.html).

Copyright © 2012 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ygeno.2012.05.007

PMID: 22609187 [Indexed for MEDLINE]

1339. J Biomol NMR. 2012 Jul;53(3):167-80. doi: 10.1007/s10858-012-9637-2. Epub 2012

Jun 8.

Resolution-by-proxy: a simple measure for assessing and comparing the overall

quality of NMR protein structures.

Berjanskii M(1), Zhou J, Liang Y, Lin G, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB, Canada.

In protein X-ray crystallography, resolution is often used as a good indicator of

structural quality. Diffraction resolution of protein crystals correlates well

with the number of X-ray observables that are used in structure generation and,

therefore, with protein coordinate errors. In protein NMR, there is no parameter

identical to X-ray resolution. Instead, resolution is often used as a synonym of

NMR model quality. Resolution of NMR structures is often deduced from ensemble

precision, torsion angle normality and number of distance restraints per residue.

The lack of common techniques to assess the resolution of X-ray and NMR

structures complicates the comparison of structures solved by these two methods.

This problem is sometimes approached by calculating "equivalent resolution" from

structure quality metrics. However, existing protocols do not offer a

comprehensive assessment of protein structure as they calculate equivalent

resolution from a relatively small number (<5) of protein parameters. Here, we

report a development of a protocol that calculates equivalent resolution from 25

measurable protein features. This new method offers better performance

(correlation coefficient of 0.92, mean absolute error of 0.28 Å) than existing

predictors of equivalent resolution. Because the method uses coordinate data as a

proxy for X-ray diffraction data, we call this measure "Resolution-by-Proxy" or

ResProx. We demonstrate that ResProx can be used to identify under-restrained,

poorly refined or inaccurate NMR structures, and can discover structural defects

that the other equivalent resolution methods cannot detect. The ResProx web

server is available at http://www.resprox.ca.

DOI: 10.1007/s10858-012-9637-2

PMID: 22678091 [Indexed for MEDLINE]

1340. J Comput Biol. 2012 Jul;19(7):887-902. doi: 10.1089/cmb.2010.0055. Epub 2011 Jan

6.

Identifying contributors of DNA mixtures by means of quantitative information of

STR typing.

Tvedebrink T(1), Eriksen PS, Mogensen HS, Morling N.

Author information:

(1)Department of Mathematical Sciences, Aalborg University, Aalborg East,

Denmark. tvede@math.aau.dk

Estimating the weight of evidence in forensic genetics is often done in terms of

a likelihood ratio, LR. The LR evaluates the probability of the observed evidence

under competing hypotheses. Most often, probabilities used in the LR only

consider the evidence from the genomic variation identified using polymorphic

genetic markers. However, modern typing techniques supply additional quantitative

data, which contain very important information about the observed evidence. This

is particularly true for cases of DNA mixtures, where more than one individual

has contributed to the observed biological stain. This article presents a method

for including the quantitative information of short tandem repeat (STR) DNA

mixtures in the LR. Also, an efficient algorithmic method for finding the best

matching combination of DNA mixture profiles is derived and implemented in an

on-line tool for two- and three-person DNA mixtures. Finally, we demonstrate for

two-person mixtures how this best matching pair of profiles can be used in

estimating the likelihood ratio using importance sampling. The reason for using

importance sampling for estimating the likelihood ratio is the often vast number

of combinations of profiles needed for the evaluation of the weight of evidence.

Online tool is available at http://people.math.aau.dk/~tvede/dna/.

DOI: 10.1089/cmb.2010.0055

PMID: 21210742 [Indexed for MEDLINE]

1341. J Med Genet. 2012 Jul;49(7):433-6. doi: 10.1136/jmedgenet-2012-100918. Epub 2012

Jun 20.

wANNOVAR: annotating genetic variants for personal genomes via the web.

Chang X(1), Wang K.

Author information:

(1)Zilkha Neurogenetic Institute, Keck School of Medicine, University of Southern

California, Los Angeles, CA, USA.

BACKGROUND: High-throughput DNA sequencing platforms have become widely

available. As a result, personal genomes are increasingly being sequenced in

research and clinical settings. However, the resulting massive amounts of

variants data pose significant challenges to the average biologists and

clinicians without bioinformatics skills.

METHODS AND RESULTS: We developed a web server called wANNOVAR to address the

critical needs for functional annotation of genetic variants from personal

genomes. The server provides simple and intuitive interface to help users

determine the functional significance of variants. These include annotating

single nucleotide variants and insertions/deletions for their effects on genes,

reporting their conservation levels (such as PhyloP and GERP++ scores),

calculating their predicted functional importance scores (such as SIFT and

PolyPhen scores), retrieving allele frequencies in public databases (such as the

1000 Genomes Project and NHLBI-ESP 5400 exomes), and implementing a 'variants

reduction' protocol to identify a subset of potentially deleterious

variants/genes. We illustrated how wANNOVAR can help draw biological insights

from sequencing data, by analysing genetic variants generated on two Mendelian

diseases.

CONCLUSIONS: We conclude that wANNOVAR will help biologists and clinicians take

advantage of the personal genome information to expedite scientific discoveries.

The wANNOVAR server is available at http://wannovar.usc.edu, and will be

continuously updated to reflect the latest annotation information.

DOI: 10.1136/jmedgenet-2012-100918

PMCID: PMC3556337

PMID: 22717648 [Indexed for MEDLINE]

1342. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W3-W12. doi: 10.1093/nar/gks632.

Epub 2012 Jun 14.

A decade of Web Server updates at the Bioinformatics Links Directory: 2003-2012.

Brazas MD(1), Yim D, Yeung W, Ouellette BF.

Author information:

(1)Ontario Institute for Cancer Research, 101 College St., Suite 800, Toronto,

Ontario, Canada M5G 0A3.

The 2012 Bioinformatics Links Directory update marks the 10th special Web Server

issue from Nucleic Acids Research. Beginning with content from their 2003

publication, the Bioinformatics Links Directory in collaboration with Nucleic

Acids Research has compiled and published a comprehensive list of freely

accessible, online tools, databases and resource materials for the bioinformatics

and life science research communities. The past decade has exhibited significant

growth and change in the types of tools, databases and resources being put forth,

reflecting both technology changes and the nature of research over that time.

With the addition of 90 web server tools and 12 updates from the July 2012 Web

Server issue of Nucleic Acids Research, the Bioinformatics Links Directory at

http://bioinformatics.ca/links\_directory/ now contains an impressive 134

resources, 455 databases and 1205 web server tools, mirroring the continued

activity and efforts of our field.

DOI: 10.1093/nar/gks632

PMCID: PMC3394264

PMID: 22700703 [Indexed for MEDLINE]

1343. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W358-63. doi:

10.1093/nar/gks577. Epub 2012 Jun 13.

Monte Carlo simulations of peptide-membrane interactions with the MCPep web

server.

Gofman Y(1), Haliloglu T, Ben-Tal N.

Author information:

(1)Helmholtz-Zentrum, Department of Structure Research on Macromolecules, 21502

Geesthacht, Germany.

The MCPep server (http://bental.tau.ac.il/MCPep/) is designed for non-experts

wishing to perform Monte Carlo (MC) simulations of helical peptides in

association with lipid membranes. MCPep is a web implementation of a previously

developed MC simulation model. The model has been tested on a variety of peptides

and protein fragments. The simulations successfully reproduced available

empirical data and provided new molecular insights, such as the preferred

locations of peptides in the membrane and the contribution of individual amino

acids to membrane association. MCPep simulates the peptide in the aqueous phase

and membrane environments, both described implicitly. In the former, the peptide

is subjected solely to internal conformational changes, and in the latter, each

MC cycle includes additional external rigid body rotational and translational

motions to allow the peptide to change its location in the membrane. The server

can explore the interaction of helical peptides of any amino-acid composition

with membranes of various lipid compositions. Given the peptide's sequence or

structure and the natural width and surface charge of the membrane, MCPep reports

the main determinants of peptide-membrane interactions, e.g. average location and

orientation in the membrane, free energy of membrane association and the

peptide's helical content. Snapshots of example simulations are also provided.

DOI: 10.1093/nar/gks577

PMCID: PMC3394254

PMID: 22695797 [Indexed for MEDLINE]

1344. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W569-72. doi:

10.1093/nar/gks576. Epub 2012 Jun 13.

EvolView, an online tool for visualizing, annotating and managing phylogenetic

trees.

Zhang H(1), Gao S, Lercher MJ, Hu S, Chen WH.

Author information:

(1)CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of

Genomics (BIG), Chinese Academy of Sciences (CAS), No.7 Beitucheng West Road,

Chaoyang District, 100029 Beijing, PR China.

EvolView is a web application for visualizing, annotating and managing

phylogenetic trees. First, EvolView is a phylogenetic tree viewer and

customization tool; it visualizes trees in various formats, customizes them

through built-in functions that can link information from external datasets, and

exports the customized results to publication-ready figures. Second, EvolView is

a tree and dataset management tool: users can easily organize related trees into

distinct projects, add new datasets to trees and edit and manage existing trees

and datasets. To make EvolView easy to use, it is equipped with an intuitive user

interface. With a free account, users can save data and manipulations on the

EvolView server. EvolView is freely available at:

http://www.evolgenius.info/evolview.html.

DOI: 10.1093/nar/gks576

PMCID: PMC3394307

PMID: 22695796 [Indexed for MEDLINE]

1345. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W228-31. doi:

10.1093/nar/gks592. Epub 2012 Jun 12.

pKNOT v.2: the protein KNOT web server.

Lai YL(1), Chen CC, Hwang JK.

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu 30068, Taiwan, Republic of China.

Knotted proteins have recently received lots of attention due to their

interesting topological novelty as well as its puzzling folding mechanisms. We

previously published a pKNOT server, which provides a structural database of

knotted proteins, analysis tools for detecting and analyzing knotted regions from

structures as well as a Java-based 3D graphics viewer for visualizing knotted

structures. However, there lacks a convenient platform performing similar tasks

directly from 'protein sequences'. In the current version of the web server,

referred to as pKNOT v.2, we implement a homology modeling tool such that the

server can now accept protein sequences in addition to 3D structures or Protein

Data Bank (PDB) IDs and return knot analysis. In addition, we have updated the

database of knotted proteins from the current PDB with a combination of automatic

and manual procedure. We believe that the updated pKNOT server with its extended

functionalities will provide better service to biologists interested in the

research of knotted proteins. The pKNOT v.2 is available from

http://pknot.life.nctu.edu.tw/.

DOI: 10.1093/nar/gks592

PMCID: PMC3394322

PMID: 22693223 [Indexed for MEDLINE]

1346. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W104-9. doi: 10.1093/nar/gks602.

Epub 2012 Jun 12.

The XXmotif web server for eXhaustive, weight matriX-based motif discovery in

nucleotide sequences.

Luehr S(1), Hartmann H, Söding J.

Author information:

(1)Gene Center, Department of Biochemistry, and Center for Integrated Protein

Science Munich (CIPSM), Ludwig-Maximilians-Universität (LMU) München,

Feodor-Lynen-Straße 25, 81377 Munich, Germany.

The discovery of regulatory motifs enriched in sets of DNA or RNA sequences is

fundamental to the analysis of a great variety of functional genomics

experiments. These motifs usually represent binding sites of proteins or

non-coding RNAs, which are best described by position weight matrices (PWMs). We

have recently developed XXmotif, a de novo motif discovery method that is able to

directly optimize the statistical significance of PWMs. XXmotif can also score

conservation and positional clustering of motifs. The XXmotif server provides (i)

a list of significantly overrepresented motif PWMs with web logos and E-values;

(ii) a graph with color-coded boxes indicating the positions of selected motifs

in the input sequences; (iii) a histogram of the overall positional distribution

for selected motifs and (iv) a page for each motif with all significant motif

occurrences, their P-values for enrichment, conservation and localization, their

sequence contexts and coordinates. Free access: http://xxmotif.genzentrum.lmu.de.

DOI: 10.1093/nar/gks602

PMCID: PMC3394272

PMID: 22693218 [Indexed for MEDLINE]

1347. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W257-62. doi:

10.1093/nar/gks555. Epub 2012 Jun 12.

BioShell Threader: protein homology detection based on sequence profiles and

secondary structure profiles.

Gront D(1), Blaszczyk M, Wojciechowski P, Kolinski A.

Author information:

(1)University of Warsaw, Faculty of Chemistry, Pasteura 1, 02-093 Warsaw, Poland.

dgront@chem.uw.edu.pl

The BioShell package has recently been extended with a web server for protein

homology detection based on profile-to-profile alignment (known as 1D threading).

Its aim is to assign structural templates to each domain of the query. The server

uses sequence profiles that describe observed sequence variability and secondary

structure profiles providing expected probability for a certain secondary

structure type at a given position in a protein. Three independent predictors are

used to increase the rate of successful predictions. Careful evaluation shows

that there is nearly 80% chance that the query sequence belongs to the same SCOP

family as the top scoring template. The Bioshell Threader server is freely

available at: http://www.bioshell.pl/threader/.

DOI: 10.1093/nar/gks555

PMCID: PMC3394251

PMID: 22693216 [Indexed for MEDLINE]

1348. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W180-5. doi: 10.1093/nar/gks551.

Epub 2012 Jun 11.

3DTF: a web server for predicting transcription factor PWMs using 3D

structure-based energy calculations.

Gabdoulline R(1), Eckweiler D, Kel A, Stegmaier P.

Author information:

(1)Heinrich-Heine University of Duesseldorf, Universitaetstr. 1, 40225

Duesseldorf, Germany.

We present the webserver 3D transcription factor (3DTF) to compute

position-specific weight matrices (PWMs) of transcription factors using a

knowledge-based statistical potential derived from crystallographic data on

protein-DNA complexes. Analysis of available structures that can be used to

construct PWMs shows that there are hundreds of 3D structures from which PWMs

could be derived, as well as thousands of proteins homologous to these.

Therefore, we created 3DTF, which delivers binding matrices given the

experimental or modeled protein-DNA complex. The webserver can be used by

biologists to derive novel PWMs for transcription factors lacking known binding

sites and is freely accessible at

http://www.gene-regulation.com/pub/programs/3dtf/.

DOI: 10.1093/nar/gks551

PMCID: PMC3394331

PMID: 22693215 [Indexed for MEDLINE]

1349. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W173-9. doi: 10.1093/nar/gks564.

Epub 2012 Jun 11.

DBD2BS: connecting a DNA-binding protein with its binding sites.

Chien TY(1), Lin CK, Lin CW, Weng YZ, Chen CY, Chang DT.

Author information:

(1)Department of Computer Science and Information Engineering, National Taiwan

University, Taipei 106, Taiwan.

By binding to short and highly conserved DNA sequences in genomes, DNA-binding

proteins initiate, enhance or repress biological processes. Accurately

identifying such binding sites, often represented by position weight matrices

(PWMs), is an important step in understanding the control mechanisms of cells.

When given coordinates of a DNA-binding domain (DBD) bound with DNA, a potential

function can be used to estimate the change of binding affinity after base

substitutions, where the changes can be summarized as a PWM. This technique

provides an effective alternative when the chromatin immunoprecipitation data are

unavailable for PWM inference. To facilitate the procedure of predicting PWMs

based on protein-DNA complexes or even structures of the unbound state, the web

server, DBD2BS, is presented in this study. The DBD2BS uses an atom-level

knowledge-based potential function to predict PWMs characterizing the sequences

to which the query DBD structure can bind. For unbound queries, a list of 1066

DBD-DNA complexes (including 1813 protein chains) is compiled for use as

templates for synthesizing bound structures. The DBD2BS provides users with an

easy-to-use interface for visualizing the PWMs predicted based on different

templates and the spatial relationships of the query protein, the DBDs and the

DNAs. The DBD2BS is the first attempt to predict PWMs of DBDs from unbound

structures rather than from bound ones. This approach increases the number of

existing protein structures that can be exploited when analyzing protein-DNA

interactions. In a recent study, the authors showed that the kernel adopted by

the DBD2BS can generate PWMs consistent with those obtained from the experimental

data. The use of DBD2BS to predict PWMs can be incorporated with sequence-based

methods to discover binding sites in genome-wide studies. Available at:

http://dbd2bs.csie.ntu.edu.tw/, http://dbd2bs.csbb.ntu.edu.tw/, and

http://dbd2bs.ee.ncku.edu.tw.

DOI: 10.1093/nar/gks564

PMCID: PMC3394304

PMID: 22693214 [Indexed for MEDLINE]

1350. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W232-7. doi: 10.1093/nar/gks529.

Epub 2012 Jun 11.

CPred: a web server for predicting viable circular permutations in proteins.

Lo WC(1), Wang LF, Liu YY, Dai T, Hwang JK, Lyu PC.

Author information:

(1)Institute of Bioinformatics and Structural Biology, National Tsing Hua

University, Hsinchu 30013, Taiwan.

Circular permutation (CP) is a protein structural rearrangement phenomenon,

through which nature allows structural homologs to have different locations of

termini and thus varied activities, stabilities and functional properties. It can

be applied in many fields of protein research and bioengineering. The limitation

of applying CP lies in its technical complexity, high cost and uncertainty of the

viability of the resulting protein variants. Not every position in a protein can

be used to create a viable circular permutant, but there is still a lack of

practical computational tools for evaluating the positional feasibility of CP

before costly experiments are carried out. We have previously designed a

comprehensive method for predicting viable CP cleavage sites in proteins. In this

work, we implement that method into an efficient and user-friendly web server

named CPred (CP site predictor), which is supposed to be helpful to promote

fundamental researches and biotechnological applications of CP. The CPred is

accessible at http://sarst.life.nthu.edu.tw/CPred.

DOI: 10.1093/nar/gks529

PMCID: PMC3394280

PMID: 22693212 [Indexed for MEDLINE]

1351. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W168-72. doi:

10.1093/nar/gks573. Epub 2012 Jun 11.

Inferring the regulatory network behind a gene expression experiment.

Bleda M(1), Medina I, Alonso R, De Maria A, Salavert F, Dopazo J.

Author information:

(1)Department of Bioinformatics and Genomics, Centro de Investigación Príncipe

Felipe (CIPF), Valencia, Spain.

Transcription factors (TFs) and miRNAs are the most important dynamic regulators

in the control of gene expression in multicellular organisms. These regulatory

elements play crucial roles in development, cell cycling and cell signaling, and

they have also been associated with many diseases. The Regulatory Network

Analysis Tool (RENATO) web server makes the exploration of regulatory networks

easy, enabling a better understanding of functional modularity and network

integrity under specific perturbations. RENATO is suitable for the analysis of

the result of expression profiling experiments. The program analyses lists of

genes and search for the regulators compatible with its activation or

deactivation. Tests of single enrichment or gene set enrichment allow the

selection of the subset of TFs or miRNAs significantly involved in the regulation

of the query genes. RENATO also offers an interactive advanced graphical

interface that allows exploring the regulatory network found.RENATO is available

at: http://renato.bioinfo.cipf.es/.

DOI: 10.1093/nar/gks573

PMCID: PMC3394273

PMID: 22693210 [Indexed for MEDLINE]

1352. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W42-8. doi: 10.1093/nar/gks560.

Epub 2012 Jun 11.

SETTER: web server for RNA structure comparison.

Cech P(1), Svozil D, Hoksza D.

Author information:

(1)Laboratory of Informatics and Chemistry, Faculty of Chemical Technology,

Prague, Czech Republic.

The recent discoveries of regulatory non-coding RNAs changed our view of RNA as a

simple information transfer molecule. Understanding the architecture and function

of active RNA molecules requires methods for comparing and analyzing their 3D

structures. While structural alignment of short RNAs is achievable in a

reasonable amount of time, large structures represent much bigger challenge.

Here, we present the SETTER web server for the RNA structure pairwise comparison

utilizing the SETTER (SEcondary sTructure-based TERtiary Structure Similarity

Algorithm) algorithm. The SETTER method divides an RNA structure into the set of

non-overlapping structural elements called generalized secondary structure units

(GSSUs). The SETTER algorithm scales as O(n(2)) with the size of a GSSUs and as

O(n) with the number of GSSUs in the structure. This scaling gives SETTER its

high speed as the average size of the GSSU remains constant irrespective of the

size of the structure. However, the favorable speed of the algorithm does not

compromise its accuracy. The SETTER web server together with the stand-alone

implementation of the SETTER algorithm are freely accessible at

http://siret.cz/setter.

DOI: 10.1093/nar/gks560

PMCID: PMC3394248

PMID: 22693209 [Indexed for MEDLINE]

1353. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W452-7. doi: 10.1093/nar/gks539.

Epub 2012 Jun 11.

SIFT web server: predicting effects of amino acid substitutions on proteins.

Sim NL(1), Kumar P, Hu J, Henikoff S, Schneider G, Ng PC.

Author information:

(1)Computational and Systems Biology, Genome Institute of Singapore, Singapore.

The Sorting Intolerant from Tolerant (SIFT) algorithm predicts the effect of

coding variants on protein function. It was first introduced in 2001, with a

corresponding website that provides users with predictions on their variants.

Since its release, SIFT has become one of the standard tools for characterizing

missense variation. We have updated SIFT's genome-wide prediction tool since our

last publication in 2009, and added new features to the insertion/deletion

(indel) tool. We also show accuracy metrics on independent data sets. The

original developers have hosted the SIFT web server at FHCRC, JCVI and the web

server is currently located at BII. The URL is http://sift-dna.org (24 May 2012,

date last accessed).

DOI: 10.1093/nar/gks539

PMCID: PMC3394338

PMID: 22689647 [Indexed for MEDLINE]

1354. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W242-8. doi: 10.1093/nar/gks559.

Epub 2012 Jun 11.

SPEER-SERVER: a web server for prediction of protein specificity determining

sites.

Chakraborty A(1), Mandloi S, Lanczycki CJ, Panchenko AR, Chakrabarti S.

Author information:

(1)Structural Biology and Bioinformatics Division, Council for Scientific and

Industrial Research (CSIR)-Indian Institute of Chemical Biology (IICB), Kolkata,

West Bengal 700032, India.

Sites that show specific conservation patterns within subsets of proteins in a

protein family are likely to be involved in the development of functional

specificity. These sites, generally termed specificity determining sites (SDS),

might play a crucial role in binding to a specific substrate or proteins.

Identification of SDS through experimental techniques is a slow, difficult and

tedious job. Hence, it is very important to develop efficient computational

methods that can more expediently identify SDS. Herein, we present Specificity

prediction using amino acids' Properties, Entropy and Evolution Rate

(SPEER)-SERVER, a web server that predicts SDS by analyzing quantitative measures

of the conservation patterns of protein sites based on their physico-chemical

properties and the heterogeneity of evolutionary changes between and within the

protein subfamilies. This web server provides an improved representation of

results, adds useful input and output options and integrates a wide range of

analysis and data visualization tools when compared with the original standalone

version of the SPEER algorithm. Extensive benchmarking finds that SPEER-SERVER

exhibits sensitivity and precision performance that, on average, meets or exceeds

that of other currently available methods. SPEER-SERVER is available at

http://www.hpppi.iicb.res.in/ss/.

DOI: 10.1093/nar/gks559

PMCID: PMC3394334

PMID: 22689646 [Indexed for MEDLINE]

1355. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W428-34. doi:

10.1093/nar/gks527. Epub 2012 Jun 11.

PBSword: a web server for searching similar protein-protein binding sites.

Pang B(1), Kuang X, Zhao N, Korkin D, Shyu CR.

Author information:

(1)Informatics Institute and Department of Computer Science, University of

Missouri, Columbia, MO, USA.

PBSword is a web server designed for efficient and accurate comparisons and

searches of geometrically similar protein-protein binding sites from a

large-scale database. The basic idea of PBSword is that each protein binding site

is first represented by a high-dimensional vector of 'visual words', which

characterizes both the global and local shape features of the binding site. It

then uses a scalable indexing technique to search for those binding sites whose

visual words representations are similar to that of the query binding site. Our

system is able to return ranked results of binding sites in short time from a

database of 194 322 domain-domain binding sites. PBSword supports query by

protein ID and by new structures uploaded by users. PBSword is a useful tool to

investigate functional connections among proteins based on the local structures

of binding site and has potential applications to protein-protein docking and

drug discovery. The system is hosted at http://pbs.rnet.missouri.edu.

DOI: 10.1093/nar/gks527

PMCID: PMC3394332

PMID: 22689645 [Indexed for MEDLINE]

1356. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W205-8. doi: 10.1093/nar/gks552.

Epub 2012 Jun 11.

MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer

specificity.

Qu W(1), Zhou Y, Zhang Y, Lu Y, Wang X, Zhao D, Yang Y, Zhang C.

Author information:

(1)Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics,

Cognitive and Mental Health Research Center, Beijing 100850, China.

Evaluating the specificity of polymerase chain reaction (PCR) primers is an

essential step in PCR primer design. The MFEprimer-2.0 server allows users to

check primer specificity against genomic DNA and messenger RNA/complementary DNA

sequence databases quickly and easily. MFEprimer-2.0 uses a k-mer index algorithm

to accelerate the search process for primer binding sites and uses thermodynamics

to evaluate binding stability between each primer and its DNA template. Several

important characteristics, such as the sequence, melting temperature and size of

each amplicon, either specific or non-specific, are reported on the results page.

Based on these characteristics and the user-friendly output, users can readily

draw conclusions about the specificity of PCR primers. Analyses for degenerate

primers and multiple PCR primers are also supported in MFEprimer-2.0. In

addition, the databases supported by MFEprimer-2.0 are comprehensive, and custom

databases can also be supported on request. The MFEprimer-2.0 server does not

require a login and is freely available at

http://biocompute.bmi.ac.cn/CZlab/MFEprimer-2.0. More over, the MFEprimer-2.0

command-line version and local server version are open source and can be

downloaded at https://github.com/quwubin/MFEprimer/wiki/Manual/.

DOI: 10.1093/nar/gks552

PMCID: PMC3394324

PMID: 22689644 [Indexed for MEDLINE]

1357. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W263-70. doi:

10.1093/nar/gks541. Epub 2012 Jun 11.

MoNetFamily: a web server to infer homologous modules and module-module

interaction networks in vertebrates.

Lin CY(1), Lin YW, Yu SW, Lo YS, Yang JM.

Author information:

(1)Institute of Bioinformatics and Systems Biology and Department of Biological

Science and Technology, National Chiao Tung University, Hsinchu, 300, Taiwan.

A module is a fundamental unit forming with highly connected proteins and

performs a certain kind of biological functions. Modules and module-module

interaction (MMI) network are essential for understanding cellular processes and

functions. The MoNetFamily web server can identify the modules, homologous

modules (called module family) and MMI networks across multiple species for the

query protein(s). This server first finds module candidates of the query by using

BLASTP to search the module template database (1785 experimental and 1252

structural templates). MoNetFamily then infers the homologous modules of the

selected module candidate using protein-protein interaction (PPI) families.

According to homologous modules and PPIs, we statistically calculated MMIs and

MMI networks across multiple species. For each module candidate, MoNetFamily

identifies its neighboring modules and their MMIs in module networks of Homo

sapiens, Mus musculus and Danio rerio. Finally, MoNetFamily shows the conserved

proteins, PPI profiles and functional annotations of the module family. Our

results indicate that the server can be useful for MMI network (e.g. 1818 modules

and 9678 MMIs in H. sapiens) visualizations and query annotations using module

families and neighboring modules. We believe that the server is able to provide

valuable insights to determine homologous modules and MMI networks across

multiple species for studying module evolution and cellular processes. The

MoNetFamily sever is available at http://monetfamily.life.nctu.edu.tw.

DOI: 10.1093/nar/gks541

PMCID: PMC3394321

PMID: 22689643 [Indexed for MEDLINE]

1358. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W147-51. doi:

10.1093/nar/gks553. Epub 2012 Jun 11.

BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction

inference.

Garcia-Garcia J(1), Schleker S, Klein-Seetharaman J, Oliva B.

Author information:

(1)Structural Bioinformatics Laboratory (GRIB-IMIM), Universitat Pompeu Fabra,

Barcelona Research Park of Biomedicine (PRBB), 08003 Barcelona, Catalonia, Spain.

Protein-protein interactions (PPIs) play a crucial role in biology, and

high-throughput experiments have greatly increased the coverage of known

interactions. Still, identification of complete inter- and intraspecies

interactomes is far from being complete. Experimental data can be complemented by

the prediction of PPIs within an organism or between two organisms based on the

known interactions of the orthologous genes of other organisms (interologs).

Here, we present the BIANA (Biologic Interactions and Network Analysis) Interolog

Prediction Server (BIPS), which offers a web-based interface to facilitate PPI

predictions based on interolog information. BIPS benefits from the capabilities

of the framework BIANA to integrate the several PPI-related databases. Additional

metadata can be used to improve the reliability of the predicted interactions.

Sensitivity and specificity of the server have been calculated using known PPIs

from different interactomes using a leave-one-out approach. The specificity is

between 72 and 98%, whereas sensitivity varies between 1 and 59%, depending on

the sequence identity cut-off used to calculate similarities between sequences.

BIPS is freely accessible at http://sbi.imim.es/BIPS.php.

DOI: 10.1093/nar/gks553

PMCID: PMC3394316

PMID: 22689642 [Indexed for MEDLINE]

1359. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W317-22. doi:

10.1093/nar/gks482. Epub 2012 Jun 11.

PredyFlexy: flexibility and local structure prediction from sequence.

de Brevern AG(1), Bornot A, Craveur P, Etchebest C, Gelly JC.

Author information:

(1)INSERM, U665, DSIMB, Paris, France. alexandre.debrevern@univ-paris-diderot.fr

Protein structures are necessary for understanding protein function at a

molecular level. Dynamics and flexibility of protein structures are also key

elements of protein function. So, we have proposed to look at protein flexibility

using novel methods: (i) using a structural alphabet and (ii) combining classical

X-ray B-factor data and molecular dynamics simulations. First, we established a

library composed of structural prototypes (LSPs) to describe protein structure by

a limited set of recurring local structures. We developed a prediction method

that proposes structural candidates in terms of LSPs and predict protein

flexibility along a given sequence. Second, we examine flexibility according to

two different descriptors: X-ray B-factors considered as good indicators of

flexibility and the root mean square fluctuations, based on molecular dynamics

simulations. We then define three flexibility classes and propose a method based

on the LSP prediction method for predicting flexibility along the sequence. This

method does not resort to sophisticate learning of flexibility but predicts

flexibility from average flexibility of predicted local structures. The method is

implemented in PredyFlexy web server. Results are similar to those obtained with

the most recent, cutting-edge methods based on direct learning of flexibility

data conducted with sophisticated algorithms. PredyFlexy can be accessed at

http://www.dsimb.inserm.fr/dsimb\_tools/predyflexy/.

DOI: 10.1093/nar/gks482

PMCID: PMC3394303

PMID: 22689641 [Indexed for MEDLINE]

1360. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W440-4. doi: 10.1093/nar/gks535.

Epub 2012 Jun 11.

PRince: a web server for structural and physicochemical analysis of protein-RNA

interface.

Barik A(1), Mishra A, Bahadur RP.

Author information:

(1)Department of Biotechnology, Indian Institute of Technology, Kharagpur 721302,

India.

We have developed a web server, PRince, which analyzes the structural features

and physicochemical properties of the protein-RNA interface. Users need to submit

a PDB file containing the atomic coordinates of both the protein and the RNA

molecules in complex form (in '.pdb' format). They should also mention the chain

identifiers of interacting protein and RNA molecules. The size of the protein-RNA

interface is estimated by measuring the solvent accessible surface area buried in

contact. For a given protein-RNA complex, PRince calculates structural,

physicochemical and hydration properties of the interacting surfaces. All these

parameters generated by the server are presented in a tabular format. The

interacting surfaces can also be visualized with software plug-in like Jmol. In

addition, the output files containing the list of the atomic coordinates of the

interacting protein, RNA and interface water molecules can be downloaded. The

parameters generated by PRince are novel, and users can correlate them with the

experimentally determined biophysical and biochemical parameters for better

understanding the specificity of the protein-RNA recognition process. This server

will be continuously upgraded to include more parameters. PRince is publicly

accessible and free for use. Available at

http://www.facweb.iitkgp.ernet.in/~rbahadur/prince/home.html.

DOI: 10.1093/nar/gks535

PMCID: PMC3394290

PMID: 22689640 [Indexed for MEDLINE]

1361. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W49-53. doi: 10.1093/nar/gks491.

Epub 2012 Jun 11.

CARNA--alignment of RNA structure ensembles.

Sorescu DA(1), Möhl M, Mann M, Backofen R, Will S.

Author information:

(1)Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, 79110

Freiburg, Germany.

Due to recent algorithmic progress, tools for the gold standard of comparative

RNA analysis, namely Sankoff-style simultaneous alignment and folding, are now

readily applicable. Such approaches, however, compare RNAs with respect to a

simultaneously predicted, single, nested consensus structure. To make multiple

alignment of RNAs available in cases, where this limitation of the standard

approach is critical, we introduce a web server that provides a complete and

convenient interface to the RNA structure alignment tool 'CARNA'. This tool

uniquely supports RNAs with multiple conserved structures per RNA and aligns

pseudoknots intrinsically; these features are highly desirable for aligning

riboswitches, RNAs with conserved folding pathways, or pseudoknots. We represent

structural input and output information as base pair probability dot plots; this

provides large flexibility in the input, ranging from fixed structures to

structure ensembles, and enables immediate visual analysis of the results. In

contrast to conventional Sankoff-style approaches, 'CARNA' optimizes all

structural similarities in the input simultaneously, for example across an entire

RNA structure ensemble. Even compared with already costly Sankoff-style

alignment, 'CARNA' solves an intrinsically much harder problem by applying

advanced, constraint-based, algorithmic techniques. Although 'CARNA' is

specialized to the alignment of RNAs with several conserved structures, its

performance on RNAs in general is on par with state-of-the-art general-purpose

RNA alignment tools, as we show in a Bralibase 2.1 benchmark. The web server is

freely available at http://rna.informatik.uni-freiburg.de/CARNA.

DOI: 10.1093/nar/gks491

PMCID: PMC3394245

PMID: 22689637 [Indexed for MEDLINE]

1362. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W484-90. doi:

10.1093/nar/gks458. Epub 2012 Jun 7.

IMP: a multi-species functional genomics portal for integration, visualization

and prediction of protein functions and networks.

Wong AK(1), Park CY, Greene CS, Bongo LA, Guan Y, Troyanskaya OG.

Author information:

(1)Department of Computer Science, Lewis-Sigler Institute for Integrative

Genomics, Princeton University, Princeton, NJ 08540, USA.

Integrative multi-species prediction (IMP) is an interactive web server that

enables molecular biologists to interpret experimental results and to generate

hypotheses in the context of a large cross-organism compendium of functional

predictions and networks. The system provides a framework for biologists to

analyze their candidate gene sets in the context of functional networks, as they

expand or focus these sets by mining functional relationships predicted from

integrated high-throughput data. IMP integrates prior knowledge and data

collections from multiple organisms in its analyses. Through flexible and

interactive visualizations, researchers can compare functional contexts and

interpret the behavior of their gene sets across organisms. Additionally, IMP

identifies homologs with conserved functional roles for knowledge transfer,

allowing for accurate function predictions even for biological processes that

have very few experimental annotations in a given organism. IMP currently

supports seven organisms (Homo sapiens, Mus musculus, Rattus novegicus,

Drosophila melanogaster, Danio rerio, Caenorhabditis elegans and Saccharomyces

cerevisiae), does not require any registration or installation and is freely

available for use at http://imp.princeton.edu.

DOI: 10.1093/nar/gks458

PMCID: PMC3394282

PMID: 22684505 [Indexed for MEDLINE]

1363. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W573-9. doi: 10.1093/nar/gks485.

Epub 2012 Jun 6.

T-REX: a web server for inferring, validating and visualizing phylogenetic trees

and networks.

Boc A(1), Diallo AB, Makarenkov V.

Author information:

(1)Département de sciences biologiques, Université de Montréal, C.P. 6128, Succ.

Centre-ville, Montréal, QC, H3C 3J7, Canada.

T-REX (Tree and reticulogram REConstruction) is a web server dedicated to the

reconstruction of phylogenetic trees, reticulation networks and to the inference

of horizontal gene transfer (HGT) events. T-REX includes several popular

bioinformatics applications such as MUSCLE, MAFFT, Neighbor Joining, NINJA,

BioNJ, PhyML, RAxML, random phylogenetic tree generator and some well-known

sequence-to-distance transformation models. It also comprises fast and effective

methods for inferring phylogenetic trees from complete and incomplete distance

matrices as well as for reconstructing reticulograms and HGT networks, including

the detection and validation of complete and partial gene transfers, inference of

consensus HGT scenarios and interactive HGT identification, developed by the

authors. The included methods allows for validating and visualizing phylogenetic

trees and networks which can be built from distance or sequence data. The web

server is available at: www.trex.uqam.ca.

DOI: 10.1093/nar/gks485

PMCID: PMC3394261

PMID: 22675075 [Indexed for MEDLINE]

1364. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W521-4. doi: 10.1093/nar/gks480.

Epub 2012 Jun 6.

Paratome: an online tool for systematic identification of antigen-binding regions

in antibodies based on sequence or structure.

Kunik V(1), Ashkenazi S, Ofran Y.

Author information:

(1)The Goodman Faculty of Life Sciences, Nanotechnology Building, Bar-Ilan

University, Ramat Gan 52900, Israel.

Antibodies are capable of specifically recognizing and binding antigens.

Identification of the antigen-binding site, commonly dubbed paratope, is of high

importance both for medical and biological applications. To date, the

identification of antigen-binding regions (ABRs) relies on tools for the

identification of complementarity-determining regions (CDRs). However, we have

shown that up to 22% of the residues that actually bind the antigen fall outside

the traditionally defined CDRs. The Paratome web server predicts the ABRs of an

antibody, given its amino acid sequence or 3D structure. It is based on a set of

consensus regions derived from a structural alignment of a non-redundant set of

all known antibody-antigen complexes. Given a query sequence or structure, the

server identifies the regions in the query antibody that correspond to the

consensus ABRs. An independent set of antibody-antigen complexes was used to test

the server and it was shown to correctly identify at least 94% of the

antigen-binding residues. The Paratome web server is freely available at

http://www.ofranlab.org/paratome/.

DOI: 10.1093/nar/gks480

PMCID: PMC3394289

PMID: 22675071 [Indexed for MEDLINE]

1365. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W329-33. doi:

10.1093/nar/gks488. Epub 2012 Jun 4.

CMWeb: an interactive on-line tool for analysing residue-residue contacts and

contact prediction methods.

Kozma D(1), Simon I, Tusnády GE.

Author information:

(1)Institute of Enzymology, Research Centre for Natural Sciences, Hungarian

Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary.

A contact map is a 2D derivative of the 3D structure of proteins, containing

various residue-residue (RR) contacts within the structure. Contact maps can be

used for the reconstruction of structure with high accuracy and can be predicted

from the amino acid sequence. Therefore understanding the various properties of

contact maps is an important step in protein structure prediction. For

investigating basic properties of contact formation and contact clusters we set

up an integrated system called Contact Map Web Viewer, or CMWeb for short. The

server can be used to visualize contact maps, to link contacts and to show them

both in 3D structures and in multiple sequence alignments and to calculate

various statistics on contacts. Moreover, we have implemented five contact

prediction methods in the CMWeb server to visualize the predicted and real RR

contacts in one contact map. The results of other RR contact prediction methods

can be uploaded as a benchmark test onto the server as well. All of these

functionality is behind a web server, thus for using our application only a

Java-capable web browser is needed, no further program installation is required.

The CMWeb is freely accessible at http://cmweb.enzim.hu.

DOI: 10.1093/nar/gks488

PMCID: PMC3394325

PMID: 22669913 [Indexed for MEDLINE]

1366. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W531-6. doi: 10.1093/nar/gks525.

Epub 2012 Jun 4.

KOSMOS: a universal morph server for nucleic acids, proteins and their complexes.

Seo S(1), Kim MK.

Author information:

(1)Department of Nano Science and Technology and School of Mechanical

Engineering, Sungkyunkwan University, 300, Cheoncheon-dong, Jangan-gu, Suwon

440-746, Korea.

KOSMOS is the first online morph server to be able to address the structural

dynamics of DNA/RNA, proteins and even their complexes, such as ribosomes. The

key functions of KOSMOS are the harmonic and anharmonic analyses of

macromolecules. In the harmonic analysis, normal mode analysis (NMA) based on an

elastic network model (ENM) is performed, yielding vibrational modes and B-factor

calculations, which provide insight into the potential biological functions of

macromolecules based on their structural features. Anharmonic analysis involving

elastic network interpolation (ENI) is used to generate plausible transition

pathways between two given conformations by optimizing a topology-oriented cost

function that guarantees a smooth transition without steric clashes. The quality

of the computed pathways is evaluated based on their various facets, including

topology, energy cost and compatibility with the NMA results. There are also two

unique features of KOSMOS that distinguish it from other morph servers: (i) the

versatility in the coarse-graining methods and (ii) the various connection rules

in the ENM. The models enable us to analyze macromolecular dynamics with the

maximum degrees of freedom by combining a variety of ENMs from full-atom to

coarse-grained, backbone and hybrid models with one connection rule, such as

distance-cutoff, number-cutoff or chemical-cutoff. KOSMOS is available at

http://bioengineering.skku.ac.kr/kosmos.

DOI: 10.1093/nar/gks525

PMCID: PMC3394317

PMID: 22669912 [Indexed for MEDLINE]

1367. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W415-22. doi:

10.1093/nar/gks515. Epub 2012 Jun 4.

Quantum.Ligand.Dock: protein-ligand docking with quantum entanglement refinement

on a GPU system.

Kantardjiev AA(1).

Author information:

(1)Biophysical Chemistry Group, Institute of Organic Chemistry, Bulgarian Academy

of Sciences, Sofia 1113, Bulgaria. alexander.kantardjiev@gmail.com

Quantum.Ligand.Dock (protein-ligand docking with graphic processing unit (GPU)

quantum entanglement refinement on a GPU system) is an original modern method for

in silico prediction of protein-ligand interactions via high-performance docking

code. The main flavour of our approach is a combination of fast search with a

special account for overlooked physical interactions. On the one hand, we take

care of self-consistency and proton equilibria mutual effects of docking

partners. On the other hand, Quantum.Ligand.Dock is the the only docking server

offering such a subtle supplement to protein docking algorithms as quantum

entanglement contributions. The motivation for development and proposition of the

method to the community hinges upon two arguments-the fundamental importance of

quantum entanglement contribution in molecular interaction and the realistic

possibility to implement it by the availability of supercomputing power. The

implementation of sophisticated quantum methods is made possible by

parallelization at several bottlenecks on a GPU supercomputer. The

high-performance implementation will be of use for large-scale virtual screening

projects, structural bioinformatics, systems biology and fundamental research in

understanding protein-ligand recognition. The design of the interface is focused

on feasibility and ease of use. Protein and ligand molecule structures are

supposed to be submitted as atomic coordinate files in PDB format. A

customization section is offered for addition of user-specified charges, extra

ionogenic groups with intrinsic pK(a) values or fixed ions. Final predicted

complexes are ranked according to obtained scores and provided in PDB format as

well as interactive visualization in a molecular viewer. Quantum.Ligand.Dock

server can be accessed at http://87.116.85.141/LigandDock.html.

DOI: 10.1093/nar/gks515

PMCID: PMC3394274

PMID: 22669908 [Indexed for MEDLINE]

1368. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W310-6. doi: 10.1093/nar/gks478.

Epub 2012 Jun 4.

NMSim web server: integrated approach for normal mode-based geometric simulations

of biologically relevant conformational transitions in proteins.

Krüger DM(1), Ahmed A, Gohlke H.

Author information:

(1)Computational Pharmaceutical Chemistry Group, Department of Mathematics and

Natural Sciences, Heinrich-Heine-University, 40225 Düsseldorf, Germany.

The NMSim web server implements a three-step approach for multiscale modeling of

protein conformational changes. First, the protein structure is coarse-grained

using the FIRST software. Second, a rigid cluster normal-mode analysis provides

low-frequency normal modes. Third, these modes are used to extend the recently

introduced idea of constrained geometric simulations by biasing backbone motions

of the protein, whereas side chain motions are biased toward favorable rotamer

states (NMSim). The generated structures are iteratively corrected regarding

steric clashes and stereochemical constraint violations. The approach allows

performing three simulation types: unbiased exploration of conformational space;

pathway generation by a targeted simulation; and radius of gyration-guided

simulation. On a data set of proteins with experimentally observed conformational

changes, the NMSim approach has been shown to be a computationally efficient

alternative to molecular dynamics simulations for conformational sampling of

proteins. The generated conformations and pathways of conformational transitions

can serve as input to docking approaches or more sophisticated sampling

techniques. The web server output is a trajectory of generated conformations,

Jmol representations of the coarse-graining and a subset of the trajectory and

data plots of structural analyses. The NMSim webserver, accessible at

http://www.nmsim.de, is free and open to all users with no login requirement.

DOI: 10.1093/nar/gks478

PMCID: PMC3394247

PMID: 22669906 [Indexed for MEDLINE]

1369. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W580-4. doi: 10.1093/nar/gks498.

Epub 2012 May 31.

FastML: a web server for probabilistic reconstruction of ancestral sequences.

Ashkenazy H(1), Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko

T.

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, 69978 Tel Aviv, Israel.

Ancestral sequence reconstruction is essential to a variety of evolutionary

studies. Here, we present the FastML web server, a user-friendly tool for the

reconstruction of ancestral sequences. FastML implements various novel features

that differentiate it from existing tools: (i) FastML uses an indel-coding

method, in which each gap, possibly spanning multiples sites, is coded as binary

data. FastML then reconstructs ancestral indel states assuming a continuous time

Markov process. FastML provides the most likely ancestral sequences, integrating

both indels and characters; (ii) FastML accounts for uncertainty in ancestral

states: it provides not only the posterior probabilities for each character and

indel at each sequence position, but also a sample of ancestral sequences from

this posterior distribution, and a list of the k-most likely ancestral sequences;

(iii) FastML implements a large array of evolutionary models, which makes it

generic and applicable for nucleotide, protein and codon sequences; and (iv) a

graphical representation of the results is provided, including, for example, a

graphical logo of the inferred ancestral sequences. The utility of FastML is

demonstrated by reconstructing ancestral sequences of the Env protein from

various HIV-1 subtypes. FastML is freely available for all academic users and is

available online at http://fastml.tau.ac.il/.

DOI: 10.1093/nar/gks498

PMCID: PMC3394241

PMID: 22661579 [Indexed for MEDLINE]

1370. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W35-41. doi: 10.1093/nar/gks513.

Epub 2012 May 31.

NASSAM: a server to search for and annotate tertiary interactions and motifs in

three-dimensional structures of complex RNA molecules.

Hamdani HY(1), Appasamy SD, Willett P, Artymiuk PJ, Firdaus-Raih M.

Author information:

(1)School of Biosciences and Biotechnology, Faculty of Science and Technology,

Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia.

Similarities in the 3D patterns of RNA base interactions or arrangements can

provide insights into their functions and roles in stabilization of the RNA 3D

structure. Nucleic Acids Search for Substructures and Motifs (NASSAM) is a graph

theoretical program that can search for 3D patterns of base arrangements by

representing the bases as pseudo-atoms. The geometric relationship of the

pseudo-atoms to each other as a pattern can be represented as a labeled graph

where the pseudo-atoms are the graph's nodes while the edges are the

inter-pseudo-atomic distances. The input files for NASSAM are PDB formatted 3D

coordinates. This web server can be used to identify matches of base arrangement

patterns in a query structure to annotated patterns that have been reported in

the literature or that have possible functional and structural stabilization

implications. The NASSAM program is freely accessible without any login

requirement at http://mfrlab.org/grafss/nassam/.

DOI: 10.1093/nar/gks513

PMCID: PMC3394293

PMID: 22661578 [Indexed for MEDLINE]

1371. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W249-56. doi:

10.1093/nar/gks481. Epub 2012 May 31.

DR\_bind: a web server for predicting DNA-binding residues from the protein

structure based on electrostatics, evolution and geometry.

Chen YC(1), Wright JD, Lim C.

Author information:

(1)Institute of Biomedical Sciences, Genomics Research Center, Academia Sinica,

Taipei 115, Taiwan.

DR\_bind is a web server that automatically predicts DNA-binding residues, given

the respective protein structure based on (i) electrostatics, (ii) evolution and

(iii) geometry. In contrast to machine-learning methods, DR\_bind does not require

a training data set or any parameters. It predicts DNA-binding residues by

detecting a cluster of conserved, solvent-accessible residues that are

electrostatically stabilized upon mutation to Asp(-)/Glu(-). The server requires

as input the DNA-binding protein structure in PDB format and outputs a

downloadable text file of the predicted DNA-binding residues, a 3D visualization

of the predicted residues highlighted in the given protein structure, and a

downloadable PyMol script for visualization of the results. Calibration on 83 and

55 non-redundant DNA-bound and DNA-free protein structures yielded a DNA-binding

residue prediction accuracy/precision of 90/47% and 88/42%, respectively. Since

DR\_bind does not require any training using protein-DNA complex structures, it

may predict DNA-binding residues in novel structures of DNA-binding proteins

resulting from structural genomics projects with no conservation data. The

DR\_bind server is freely available with no login requirement at

http://dnasite.limlab.ibms.sinica.edu.tw.

DOI: 10.1093/nar/gks481

PMCID: PMC3394278

PMID: 22661576 [Indexed for MEDLINE]

1372. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W294-7. doi: 10.1093/nar/gks493.

Epub 2012 May 30.

GalaxyWEB server for protein structure prediction and refinement.

Ko J(1), Park H, Heo L, Seok C.

Author information:

(1)Department of Chemistry, Seoul National University, Seoul 151-747, Korea.

Three-dimensional protein structures provide invaluable information for

understanding and regulating biological functions of proteins. The GalaxyWEB

server predicts protein structure from sequence by template-based modeling and

refines loop or terminus regions by ab initio modeling. This web server is based

on the method tested in CASP9 (9th Critical Assessment of techniques for protein

Structure Prediction) as 'Seok-server', which was assessed to be among top

performing template-based modeling servers. The method generates reliable core

structures from multiple templates and re-builds unreliable loops or termini by

using an optimization-based refinement method. In addition to structure

prediction, a user can also submit a refinement only job by providing a starting

model structure and locations of loops or termini to refine. The web server can

be freely accessed at http://galaxy.seoklab.org/.

DOI: 10.1093/nar/gks493

PMCID: PMC3394311

PMID: 22649060 [Indexed for MEDLINE]

1373. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W498-504. doi:

10.1093/nar/gks494. Epub 2012 May 30.

DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in

pathways.

Vlachos IS(1), Kostoulas N, Vergoulis T, Georgakilas G, Reczko M, Maragkakis M,

Paraskevopoulou MD, Prionidis K, Dalamagas T, Hatzigeorgiou AG.

Author information:

(1)DIANA-Lab, Institute of Molecular Oncology, Biomedical Sciences Research

Center 'Alexander Fleming', 16672 Vari, Greece.

MicroRNAs (miRNAs) are key regulators of diverse biological processes and their

functional analysis has been deemed central in many research pipelines. The new

version of DIANA-miRPath web server was redesigned from the ground-up. The user

of DNA Intelligent Analysis (DIANA) DIANA-miRPath v2.0 can now utilize miRNA

targets predicted with high accuracy based on DIANA-microT-CDS and/or

experimentally verified targets from TarBase v6; combine results with merging and

meta-analysis algorithms; perform hierarchical clustering of miRNAs and pathways

based on their interaction levels; as well as elaborate sophisticated

visualizations, such as dendrograms or miRNA versus pathway heat maps, from an

intuitive and easy to use web interface. New modules enable DIANA-miRPath server

to provide information regarding pathogenic single nucleotide polymorphisms

(SNPs) in miRNA target sites (SNPs module) or to annotate all the predicted and

experimentally validated miRNA targets in a selected molecular pathway (Reverse

Search module). DIANA-miRPath v2.0 is an efficient and yet easy to use tool that

can be incorporated successfully into miRNA-related analysis pipelines. It

provides for the first time a series of highly specific tools for miRNA-targeted

pathway analysis via a web interface and can be accessed at

http://www.microrna.gr/miRPathv2.

DOI: 10.1093/nar/gks494

PMCID: PMC3394305

PMID: 22649059 [Indexed for MEDLINE]

1374. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W393-9. doi: 10.1093/nar/gks496.

Epub 2012 May 30.

idTarget: a web server for identifying protein targets of small chemical

molecules with robust scoring functions and a divide-and-conquer docking

approach.

Wang JC(1), Chu PY, Chen CM, Lin JH.

Author information:

(1)Institute of Biomedical Engineering, National Taiwan University, Taipei,

Taiwan.

Identification of possible protein targets of small chemical molecules is an

important step for unravelling their underlying causes of actions at the

molecular level. To this end, we construct a web server, idTarget, which can

predict possible binding targets of a small chemical molecule via a

divide-and-conquer docking approach, in combination with our recently developed

scoring functions based on robust regression analysis and quantum chemical charge

models. Affinity profiles of the protein targets are used to provide the

confidence levels of prediction. The divide-and-conquer docking approach uses

adaptively constructed small overlapping grids to constrain the searching space,

thereby achieving better docking efficiency. Unlike previous approaches that

screen against a specific class of targets or a limited number of targets,

idTarget screen against nearly all protein structures deposited in the Protein

Data Bank (PDB). We show that idTarget is able to reproduce known off-targets of

drugs or drug-like compounds, and the suggested new targets could be prioritized

for further investigation. idTarget is freely available as a web-based server at

http://idtarget.rcas.sinica.edu.tw.

DOI: 10.1093/nar/gks496

PMCID: PMC3394295

PMID: 22649057 [Indexed for MEDLINE]

1375. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W88-95. doi: 10.1093/nar/gks497.

Epub 2012 May 29.

METAGENassist: a comprehensive web server for comparative metagenomics.

Arndt D(1), Xia J, Liu Y, Zhou Y, Guo AC, Cruz JA, Sinelnikov I, Budwill K, Nesbø

CL, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, Alberta,

Canada.

With recent improvements in DNA sequencing and sample extraction techniques, the

quantity and quality of metagenomic data are now growing exponentially. This

abundance of richly annotated metagenomic data and bacterial census information

has spawned a new branch of microbiology called comparative metagenomics.

Comparative metagenomics involves the comparison of bacterial populations between

different environmental samples, different culture conditions or different

microbial hosts. However, in order to do comparative metagenomics, one typically

requires a sophisticated knowledge of multivariate statistics and/or advanced

software programming skills. To make comparative metagenomics more accessible to

microbiologists, we have developed a freely accessible, easy-to-use web server

for comparative metagenomic analysis called METAGENassist. Users can upload their

bacterial census data from a wide variety of common formats, using either

amplified 16S rRNA data or shotgun metagenomic data. Metadata concerning

environmental, culture, or host conditions can also be uploaded. During the data

upload process, METAGENassist also performs an automated taxonomic-to-phenotypic

mapping. Phenotypic information covering nearly 20 functional categories such as

GC content, genome size, oxygen requirements, energy sources and preferred

temperature range is automatically generated from the taxonomic input data. Using

this phenotypically enriched data, users can then perform a variety of

multivariate and univariate data analyses including fold change analysis,

t-tests, PCA, PLS-DA, clustering and classification. To facilitate data

processing, users are guided through a step-by-step analysis workflow using a

variety of menus, information hyperlinks and check boxes. METAGENassist also

generates colorful, publication quality tables and graphs that can be downloaded

and used directly in the preparation of scientific papers. METAGENassist is

available at http://www.metagenassist.ca.

DOI: 10.1093/nar/gks497

PMCID: PMC3394294

PMID: 22645318 [Indexed for MEDLINE]

1376. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W445-51. doi:

10.1093/nar/gks479. Epub 2012 May 29.

dbCAN: a web resource for automated carbohydrate-active enzyme annotation.

Yin Y(1), Mao X, Yang J, Chen X, Mao F, Xu Y.

Author information:

(1)Computational System Biology Laboratory, Department of Biochemistry and

Molecular Biology, Institute of Bioinformatics, BioEnergy Science Center,

University of Georgia, Athens, GA, USA.

Carbohydrate-active enzymes (CAZymes) are very important to the biotech industry,

particularly the emerging biofuel industry because CAZymes are responsible for

the synthesis, degradation and modification of all the carbohydrates on Earth. We

have developed a web resource, dbCAN

(http://csbl.bmb.uga.edu/dbCAN/annotate.php), to provide a capability for

automated CAZyme signature domain-based annotation for any given protein data set

(e.g. proteins from a newly sequenced genome) submitted to our server. To

accomplish this, we have explicitly defined a signature domain for every CAZyme

family, derived based on the CDD (conserved domain database) search and

literature curation. We have also constructed a hidden Markov model to represent

the signature domain of each CAZyme family. These CAZyme family-specific HMMs are

our key contribution and the foundation for the automated CAZyme annotation.

DOI: 10.1093/nar/gks479

PMCID: PMC3394287

PMID: 22645317 [Indexed for MEDLINE]

1377. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W71-5. doi: 10.1093/nar/gks474.

Epub 2012 May 27.

KD4v: Comprehensible Knowledge Discovery System for Missense Variant.

Luu TD(1), Rusu A, Walter V, Linard B, Poidevin L, Ripp R, Moulinier L, Muller J,

Raffelsberger W, Wicker N, Lecompte O, Thompson JD, Poch O, Nguyen H.

Author information:

(1)Laboratoire de Bioinformatique et Génomique Intégratives, Institut de

Génétique et de Biologie Moléculaire et Cellulaire, 67404 Illkirch, France.

A major challenge in the post-genomic era is a better understanding of how human

genetic alterations involved in disease affect the gene products. The KD4v

(Comprehensible Knowledge Discovery System for Missense Variant) server allows to

characterize and predict the phenotypic effects (deleterious/neutral) of missense

variants. The server provides a set of rules learned by Induction Logic

Programming (ILP) on a set of missense variants described by conservation,

physico-chemical, functional and 3D structure predicates. These rules are

interpretable by non-expert humans and are used to accurately predict the

deleterious/neutral status of an unknown mutation. The web server is available at

http://decrypthon.igbmc.fr/kd4v.

DOI: 10.1093/nar/gks474

PMCID: PMC3394327

PMID: 22641855 [Indexed for MEDLINE]

1378. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W466-70. doi:

10.1093/nar/gks489. Epub 2012 May 27.

CombFunc: predicting protein function using heterogeneous data sources.

Wass MN(1), Barton G, Sternberg MJ.

Author information:

(1)Centre for Bioinformatics, Imperial College London, London, SW7 2AZ, UK.

mark.wass04@imperial.ac.uk

Only a small fraction of known proteins have been functionally characterized,

making protein function prediction essential to propose annotations for

uncharacterized proteins. In recent years many function prediction methods have

been developed using various sources of biological data from protein sequence and

structure to gene expression data. Here we present the CombFunc web server, which

makes Gene Ontology (GO)-based protein function predictions. CombFunc

incorporates ConFunc, our existing function prediction method, with other

approaches for function prediction that use protein sequence, gene expression and

protein-protein interaction data. In benchmarking on a set of 1686 proteins

CombFunc obtains precision and recall of 0.71 and 0.64 respectively for gene

ontology molecular function terms. For biological process GO terms precision of

0.74 and recall of 0.41 is obtained. CombFunc is available at

http://www.sbg.bio.ic.ac.uk/combfunc.

DOI: 10.1093/nar/gks489

PMCID: PMC3394346

PMID: 22641853 [Indexed for MEDLINE]

1379. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W592-6. doi: 10.1093/nar/gks448.

Epub 2012 May 28.

GGRNA: an ultrafast, transcript-oriented search engine for genes and transcripts.

Naito Y(1), Bono H.

Author information:

(1)Database Center for Life Science, Research Organization of Information and

Systems, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan.

GGRNA (http://GGRNA.dbcls.jp/) is a Google-like, ultrafast search engine for

genes and transcripts. The web server accepts arbitrary words and phrases, such

as gene names, IDs, gene descriptions, annotations of gene and even

nucleotide/amino acid sequences through one simple search box, and quickly

returns relevant RefSeq transcripts. A typical search takes just a few seconds,

which dramatically enhances the usability of routine searching. In particular,

GGRNA can search sequences as short as 10 nt or 4 amino acids, which cannot be

handled easily by popular sequence analysis tools. Nucleotide sequences can be

searched allowing up to three mismatches, or the query sequences may contain

degenerate nucleotide codes (e.g. N, R, Y, S). Furthermore, Gene Ontology

annotations, Enzyme Commission numbers and probe sequences of catalog microarrays

are also incorporated into GGRNA, which may help users to conduct searches by

various types of keywords. GGRNA web server will provide a simple and powerful

interface for finding genes and transcripts for a wide range of users. All

services at GGRNA are provided free of charge to all users.

DOI: 10.1093/nar/gks448

PMCID: PMC3394333

PMID: 22641850 [Indexed for MEDLINE]

1380. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W334-9. doi: 10.1093/nar/gks436.

Epub 2012 May 25.

Super: a web server to rapidly screen superposable oligopeptide fragments from

the protein data bank.

Collier JH(1), Lesk AM, Garcia de la Banda M, Konagurthu AS.

Author information:

(1)Faculty of Information Technology, Monash University, Clayton, Victoria 3800,

Australia.

Searching for well-fitting 3D oligopeptide fragments within a large collection of

protein structures is an important task central to many analyses involving

protein structures. This article reports a new web server, Super, dedicated to

the task of rapidly screening the protein data bank (PDB) to identify all

fragments that superpose with a query under a prespecified threshold of

root-mean-square deviation (RMSD). Super relies on efficiently computing a

mathematical bound on the commonly used structural similarity measure, RMSD of

superposition. This allows the server to filter out a large proportion of

fragments that are unrelated to the query; >99% of the total number of fragments

in some cases. For a typical query, Super scans the current PDB containing over

80,500 structures (with ∼40 million potential oligopeptide fragments to match) in

under a minute. Super web server is freely accessible from:

http://lcb.infotech.monash.edu.au/super.

DOI: 10.1093/nar/gks436

PMCID: PMC3394326

PMID: 22638586 [Indexed for MEDLINE]

1381. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W281-7. doi: 10.1093/nar/gks469.

Epub 2012 May 25.

Seq2Logo: a method for construction and visualization of amino acid binding

motifs and sequence profiles including sequence weighting, pseudo counts and

two-sided representation of amino acid enrichment and depletion.

Thomsen MC(1), Nielsen M.

Author information:

(1)Center for Biological Sequence Analysis, Technical University of Denmark,

DK-2800 Kgs. Lyngby, Denmark.

Seq2Logo is a web-based sequence logo generator. Sequence logos are a graphical

representation of the information content stored in a multiple sequence alignment

(MSA) and provide a compact and highly intuitive representation of the

position-specific amino acid composition of binding motifs, active sites, etc. in

biological sequences. Accurate generation of sequence logos is often compromised

by sequence redundancy and low number of observations. Moreover, most methods

available for sequence logo generation focus on displaying the position-specific

enrichment of amino acids, discarding the equally valuable information related to

amino acid depletion. Seq2logo aims at resolving these issues allowing the user

to include sequence weighting to correct for data redundancy, pseudo counts to

correct for low number of observations and different logotype representations

each capturing different aspects related to amino acid enrichment and depletion.

Besides allowing input in the format of peptides and MSA, Seq2Logo accepts input

as Blast sequence profiles, providing easy access for non-expert end-users to

characterize and identify functionally conserved/variable amino acids in any

given protein of interest. The output from the server is a sequence logo and a

PSSM. Seq2Logo is available at http://www.cbs.dtu.dk/biotools/Seq2Logo (14 May

2012, date last accessed).

DOI: 10.1093/nar/gks469

PMCID: PMC3394285

PMID: 22638583 [Indexed for MEDLINE]

1382. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W199-204. doi:

10.1093/nar/gks450. Epub 2012 May 25.

AVPpred: collection and prediction of highly effective antiviral peptides.

Thakur N(1), Qureshi A, Kumar M.

Author information:

(1)Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39-A,

Chandigarh 160036, India.

In the battle against viruses, antiviral peptides (AVPs) had demonstrated the

immense potential. Presently, more than 15 peptide-based drugs are in various

stages of clinical trials. Emerging and re-emerging viruses further emphasize the

efforts to accelerate antiviral drug discovery efforts. Despite, huge importance

of the field, no dedicated AVP resource is available. In the present study, we

have collected 1245 peptides which were experimentally checked for antiviral

activity targeting important human viruses like influenza, HIV, HCV and SARS,

etc. After removing redundant peptides, 1056 peptides were divided into 951

training and 105 validation data sets. We have exploited various peptides

sequence features, i.e. motifs and alignment followed by amino acid composition

and physicochemical properties during 5-fold cross validation using Support

Vector Machine. Physiochemical properties-based model achieved maximum 85%

accuracy and 0.70 Matthew's Correlation Coefficient (MCC). Performance of this

model on the experimental validation data set showed 86% accuracy and 0.71 MCC

which is far better than the general antimicrobial peptides prediction methods.

Therefore, AVPpred-the first web server for predicting the highly effective AVPs

would certainly be helpful to researchers working on peptide-based antiviral

development. The web server is freely available at

http://crdd.osdd.net/servers/avppred.

DOI: 10.1093/nar/gks450

PMCID: PMC3394244

PMID: 22638580 [Indexed for MEDLINE]

1383. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W505-9. doi: 10.1093/nar/gks445.

Epub 2012 May 25.

SteinerNet: a web server for integrating 'omic' data to discover hidden

components of response pathways.

Tuncbag N(1), McCallum S, Huang SS, Fraenkel E.

Author information:

(1)Department of Biological Engineering, Massachusetts Institute of Technology,

Cambridge, MA 02139, USA.

High-throughput technologies including transcriptional profiling, proteomics and

reverse genetics screens provide detailed molecular descriptions of cellular

responses to perturbations. However, it is difficult to integrate these diverse

data to reconstruct biologically meaningful signaling networks. Previously, we

have established a framework for integrating transcriptional, proteomic and

interactome data by searching for the solution to the prize-collecting Steiner

tree problem. Here, we present a web server, SteinerNet, to make this method

available in a user-friendly format for a broad range of users with data from any

species. At a minimum, a user only needs to provide a set of experimentally

detected proteins and/or genes and the server will search for connections among

these data from the provided interactomes for yeast, human, mouse, Drosophila

melanogaster and Caenorhabditis elegans. More advanced users can upload their own

interactome data as well. The server provides interactive visualization of the

resulting optimal network and downloadable files detailing the analysis and

results. We believe that SteinerNet will be useful for researchers who would like

to integrate their high-throughput data for a specific condition or cellular

response and to find biologically meaningful pathways. SteinerNet is accessible

at http://fraenkel.mit.edu/steinernet.

DOI: 10.1093/nar/gks445

PMCID: PMC3394335

PMID: 22638579 [Indexed for MEDLINE]

1384. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W364-9. doi: 10.1093/nar/gks444.

Epub 2012 May 25.

iELM--a web server to explore short linear motif-mediated interactions.

Weatheritt RJ(1), Jehl P, Dinkel H, Gibson TJ.

Author information:

(1)Structural and Computational Biology Unit, European Molecular Biology

Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

The recent expansion in our knowledge of protein-protein interactions (PPIs) has

allowed the annotation and prediction of hundreds of thousands of interactions.

However, the function of many of these interactions remains elusive. The

interactions of Eukaryotic Linear Motif (iELM) web server provides a resource for

predicting the function and positional interface for a subset of interactions

mediated by short linear motifs (SLiMs). The iELM prediction algorithm is based

on the annotated SLiM classes from the Eukaryotic Linear Motif (ELM) resource and

allows users to explore both annotated and user-generated PPI networks for

SLiM-mediated interactions. By incorporating the annotated information from the

ELM resource, iELM provides functional details of PPIs. This can be used in

proteomic analysis, for example, to infer whether an interaction promotes complex

formation or degradation. Furthermore, details of the molecular interface of the

SLiM-mediated interactions are also predicted. This information is displayed in a

fully searchable table, as well as graphically with the modular architecture of

the participating proteins extracted from the UniProt and Phospho.ELM resources.

A network figure is also presented to aid the interpretation of results. The iELM

server supports single protein queries as well as large-scale proteomic

submissions and is freely available at http://i.elm.eu.org.

DOI: 10.1093/nar/gks444

PMCID: PMC3394315

PMID: 22638578 [Indexed for MEDLINE]

1385. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W547-52. doi:

10.1093/nar/gks449. Epub 2012 May 25.

DichroMatch: a website for similarity searching of circular dichroism spectra.

Klose DP(1), Wallace BA, Janes RW.

Author information:

(1)School of Biological and Chemical Sciences, Queen Mary University of London,

London E1 4NS, UK.

Circular dichroism (CD) spectroscopy is a widely used method for examining the

structure, folding and conformational changes of proteins. A new online CD

analysis server (DichroMatch) has been developed for identifying proteins with

similar spectral characteristics by detecting possible structurally and

functionally related proteins and homologues. DichroMatch includes six different

methods for determining the spectral nearest neighbours to a query protein

spectrum and provides metrics of how similar these spectra are and, if

corresponding crystal structures are available for the closest matched proteins,

information on their secondary structures and fold classifications. By default,

DichroMatch uses all the entries in the Protein Circular Dichroism Data Bank

(PCDDB) for its comparison set, providing the broadest range of publicly

available protein spectra to match with the unknown protein. Alternatively, users

can download or create their own specialized data sets, thereby enabling

comparisons between the structures of related proteins such as wild-type versus

mutants or homologues or a series of spectra of the same protein under different

conditions. The DichroMatch server is freely available at

http://dichromatch.cryst.bbk.ac.uk.

DOI: 10.1093/nar/gks449

PMCID: PMC3394267

PMID: 22638573 [Indexed for MEDLINE]

1386. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W400-8. doi: 10.1093/nar/gks421.

Epub 2012 May 22.

PocketAnnotate: towards site-based function annotation.

Anand P(1), Yeturu K, Chandra N.

Author information:

(1)Department of Biochemistry, Indian Institute of Science, Bangalore 560012,

Karnataka, India.

A computational pipeline PocketAnnotate for functional annotation of proteins at

the level of binding sites has been proposed in this study. The pipeline

integrates three in-house algorithms for site-based function annotation:

PocketDepth, for prediction of binding sites in protein structures; PocketMatch,

for rapid comparison of binding sites and PocketAlign, to obtain detailed

alignment between pair of binding sites. A novel scheme has been developed to

rapidly generate a database of non-redundant binding sites. For a given input

protein structure, putative ligand-binding sites are identified, matched in real

time against the database and the query substructure aligned with the promising

hits, to obtain a set of possible ligands that the given protein could bind to.

The input can be either whole protein structures or merely the substructures

corresponding to possible binding sites. Structure-based function annotation at

the level of binding sites thus achieved could prove very useful for cases where

no obvious functional inference can be obtained based purely on sequence or

fold-level analyses. An attempt has also been made to analyse proteins of no

known function from Protein Data Bank. PocketAnnotate would be a valuable tool

for the scientific community and contribute towards structure-based functional

inference. The web server can be freely accessed at

http://proline.biochem.iisc.ernet.in/pocketannotate/.

DOI: 10.1093/nar/gks421

PMCID: PMC3394344

PMID: 22618878 [Indexed for MEDLINE]

1387. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W82-7. doi: 10.1093/nar/gks418.

Epub 2012 May 22.

TaxMan: a server to trim rRNA reference databases and inspect taxonomic coverage.

Brandt BW(1), Bonder MJ, Huse SM, Zaura E.

Author information:

(1)Department of Preventive Dentistry, Academic Centre for Dentistry Amsterdam

(ACTA), University of Amsterdam and VU University Amsterdam, Amsterdam, The

Netherlands. b.brandt@acta.nl

Amplicon sequencing of the hypervariable regions of the small subunit ribosomal

RNA gene is a widely accepted method for identifying the members of complex

bacterial communities. Several rRNA gene sequence reference databases can be used

to assign taxonomic names to the sequencing reads using BLAST, USEARCH, GAST or

the RDP classifier. Next-generation sequencing methods produce ample reads, but

they are short, currently ∼100-450 nt (depending on the technology), as compared

to the full rRNA gene of ∼1550 nt. It is important, therefore, to select the

right rRNA gene region for sequencing. The primers should amplify the species of

interest and the hypervariable regions should differentiate their taxonomy. Here,

we introduce TaxMan: a web-based tool that trims reference sequences based on

user-selected primer pairs and returns an assessment of the primer specificity by

taxa. It allows interactive plotting of taxa, both amplified and missed in silico

by the primers used. Additionally, using the trimmed sequences improves the speed

of sequence matching algorithms. The smaller database greatly improves run times

(up to 98%) and memory usage, not only of similarity searching (BLAST), but also

of chimera checking (UCHIME) and of clustering the reads (UCLUST). TaxMan is

available at http://www.ibi.vu.nl/programs/taxmanwww/.

DOI: 10.1093/nar/gks418

PMCID: PMC3394339

PMID: 22618877 [Indexed for MEDLINE]

1388. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W157-61. doi:

10.1093/nar/gks446. Epub 2012 May 22.

NetAligner--a network alignment server to compare complexes, pathways and whole

interactomes.

Pache RA(1), Céol A, Aloy P.

Author information:

(1)Institute for Research in Biomedicine (IRB) Barcelona, Department of

Structural and Computational Biology, Joint IRB-BSC Program in Computational

Biology, c/Baldiri Reixac 10-12, 08028 Barcelona, Spain.

The many ongoing genome sequencing initiatives are delivering comprehensive lists

of the individual molecular components present in an organism, but these reveal

little about how they work together. Follow-up initiatives are revealing

thousands of interrelationships between gene products that need to be analyzed

with novel bioinformatics approaches able to capture their complex emerging

properties. Recently, we developed NetAligner, a novel network alignment tool

that allows the identification of conserved protein complexes and pathways across

organisms, providing valuable hints as to how those interaction networks evolved.

NetAligner includes the prediction of likely conserved interactions, based on

evolutionary distances, to counter the high number of missing interactions in

current interactome networks, and a fast assessment of the statistical

significance of individual alignment solutions, which increases its performance

with respect to existing tools. The web server implementation of the NetAligner

algorithm presented here features complex, pathway and interactome to interactome

alignments for seven model organisms, namely Homo sapiens, Mus musculus,

Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana,

Saccharomyces cerevisiae and Escherichia coli. The user can query complexes and

pathways of arbitrary topology against a target species interactome, or directly

compare two complete interactomes to identify conserved complexes and

subnetworks. Alignment solutions can be downloaded or directly visualized in the

browser. The NetAligner web server is publicly available at

http://netaligner.irbbarcelona.org/.

DOI: 10.1093/nar/gks446

PMCID: PMC3394252

PMID: 22618871 [Indexed for MEDLINE]

1389. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W76-81. doi: 10.1093/nar/gks397.

Epub 2012 May 22.

VarioWatch: providing large-scale and comprehensive annotations on human genomic

variants in the next generation sequencing era.

Cheng YC(1), Hsiao FC, Yeh EC, Lin WJ, Tang CY, Tseng HC, Wu HT, Liu CK, Chen CC,

Chen YT, Yao A.

Author information:

(1)National Center for Genome Medicine and Institute of Biomedical Sciences,

Academia Sinica, Taiwan 11529, R.O.C.

VarioWatch (http://genepipe.ncgm.sinica.edu.tw/variowatch/) has been vastly

improved since its former publication GenoWatch in the 2008 Web Server Issue. It

is now at least 10 000-times faster in annotating a variant. Drastic speed

increase, through complete re-design of its working mechanism, makes VarioWatch

capable of annotating millions of human genomic variants generated from next

generation sequencing in minutes, if not seconds. While using MegaQuery of

VarioWatch to quickly annotate variants, users can apply various filters to

retrieve a subgroup of variants according to the risk levels, interested regions,

etc. that satisfy users' requirements. In addition to performance leap, many new

features have also been added, such as annotation on novel variants, functional

analyses on splice sites and in/dels, detailed variant information in tabulated

form, plus a risk level decision tree regarding the analyzed variant. Up to 1000

target variants can be visualized with our carefully designed Genome View, Gene

View, Transcript View and Variation View. Two commonly used reference versions,

NCBI build 36.3 and NCBI build 37.2, are supported. VarioWatch is unique in its

ability to annotate comprehensively and efficiently millions of variants online,

immediately delivering the results in real time, plus visualizes up to 1000

annotated variants.

DOI: 10.1093/nar/gks397

PMCID: PMC3394242

PMID: 22618869 [Indexed for MEDLINE]

1390. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W162-7. doi: 10.1093/nar/gks459.

Epub 2012 May 18.

GENIES: gene network inference engine based on supervised analysis.

Kotera M(1), Yamanishi Y, Moriya Y, Kanehisa M, Goto S.

Author information:

(1)Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji,

Kyoto 611-0011, Japan.

Gene network inference engine based on supervised analysis (GENIES) is a web

server to predict unknown part of gene network from various types of genome-wide

data in the framework of supervised network inference. The originality of GENIES

lies in the construction of a predictive model using partially known network

information and in the integration of heterogeneous data with kernel methods. The

GENIES server accepts any 'profiles' of genes or proteins (e.g. gene expression

profiles, protein subcellular localization profiles and phylogenetic profiles) or

pre-calculated gene-gene similarity matrices (or 'kernels') in the tab-delimited

file format. As a training data set to learn a predictive model, the users can

choose either known molecular network information in the KEGG PATHWAY database or

their own gene network data. The user can also select an algorithm of supervised

network inference, choose various parameters in the method, and control the

weights of heterogeneous data integration. The server provides the list of newly

predicted gene pairs, maps the predicted gene pairs onto the associated pathway

diagrams in KEGG PATHWAY and indicates candidate genes for missing enzymes in

organism-specific metabolic pathways. GENIES (http://www.genome.jp/tools/genies/)

is publicly available as one of the genome analysis tools in GenomeNet.

DOI: 10.1093/nar/gks459

PMCID: PMC3394336

PMID: 22610856 [Indexed for MEDLINE]

1391. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W553-9. doi: 10.1093/nar/gks420.

Epub 2012 May 16.

Cyber-T web server: differential analysis of high-throughput data.

Kayala MA(1), Baldi P.

Author information:

(1)Department of Computer Science and Institute for Genomics and Bioinformatics,

University of California, Irvine, CA 92697, USA.

The Bayesian regularization method for high-throughput differential analysis,

described in Baldi and Long (A Bayesian framework for the analysis of microarray

expression data: regularized t-test and statistical inferences of gene changes.

Bioinformatics 2001: 17: 509-519) and implemented in the Cyber-T web server, is

one of the most widely validated. Cyber-T implements a t-test using a Bayesian

framework to compute a regularized variance of the measurements associated with

each probe under each condition. This regularized estimate is derived by flexibly

combining the empirical measurements with a prior, or background, derived from

pooling measurements associated with probes in the same neighborhood. This

approach flexibly addresses problems associated with low replication levels and

technology biases, not only for DNA microarrays, but also for other technologies,

such as protein arrays, quantitative mass spectrometry and next-generation

sequencing (RNA-seq). Here we present an update to the Cyber-T web server,

incorporating several useful new additions and improvements. Several

preprocessing data normalization options including logarithmic and (Variance

Stabilizing Normalization) VSN transforms are included. To augment two-sample

t-tests, a one-way analysis of variance is implemented. Several methods for

multiple tests correction, including standard frequentist methods and a

probabilistic mixture model treatment, are available. Diagnostic plots allow

visual assessment of the results. The web server provides comprehensive

documentation and example data sets. The Cyber-T web server, with R source code

and data sets, is publicly available at http://cybert.ics.uci.edu/.

DOI: 10.1093/nar/gks420

PMCID: PMC3394347

PMID: 22600740 [Indexed for MEDLINE]

1392. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W193-8. doi: 10.1093/nar/gks414.

Epub 2012 May 16.

jpHMM: recombination analysis in viruses with circular genomes such as the

hepatitis B virus.

Schultz AK(1), Bulla I, Abdou-Chekaraou M, Gordien E, Morgenstern B, Zoaulim F,

Dény P, Stanke M.

Author information:

(1)Abteilung für Bioinformatik, Institut für Mikrobiologie und Genetik,

Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany.

anne@gobics.de

jpHMM is a very accurate and widely used tool for recombination detection in

genomic sequences of HIV-1. Here, we present an extension of jpHMM to analyze

recombinations in viruses with circular genomes such as the hepatitis B virus

(HBV). Sequence analysis of circular genomes is usually performed on linearized

sequences using linear models. Since linear models are unable to model

dependencies between nucleotides at the 5'- and 3'-end of a sequence, this can

result in inaccurate predictions of recombination breakpoints and thus in

incorrect classification of viruses with circular genomes. The proposed circular

jpHMM takes into account the circularity of the genome and is not biased against

recombination breakpoints close to the 5'- or 3'-end of the linearized version of

the circular genome. It can be applied automatically to any query sequence

without assuming a specific origin for the sequence coordinates. We apply the

method to genomic sequences of HBV and visualize its output in a circular form.

jpHMM is available online at http://jphmm.gobics.de for download and as a web

server for HIV-1 and HBV sequences.

DOI: 10.1093/nar/gks414

PMCID: PMC3394342

PMID: 22600739 [Indexed for MEDLINE]

1393. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W423-7. doi: 10.1093/nar/gks398.

Epub 2012 May 16.

PepSite: prediction of peptide-binding sites from protein surfaces.

Trabuco LG(1), Lise S, Petsalaki E, Russell RB.

Author information:

(1)CellNetworks, University of Heidelberg, 69120 Heidelberg, Germany.

Complex biological functions emerge through intricate protein-protein interaction

networks. An important class of protein-protein interaction corresponds to

peptide-mediated interactions, in which a short peptide stretch from one partner

interacts with a large protein surface from the other partner. Protein-peptide

interactions are typically of low affinity and involved in regulatory mechanisms,

dynamically reshaping protein interaction networks. Due to the relatively small

interaction surface, modulation of protein-peptide interactions is feasible and

highly attractive for therapeutic purposes. Unfortunately, the number of

available 3D structures of protein-peptide interfaces is very limited. For

typical cases where a protein-peptide structure of interest is not available, the

PepSite web server can be used to predict peptide-binding spots from protein

surfaces alone. The PepSite method relies on preferred peptide-binding

environments calculated from a set of known protein-peptide 3D structures,

combined with distance constraints derived from known peptides. We present an

updated version of the web server that is orders of magnitude faster than the

original implementation, returning results in seconds instead of minutes or

hours. The PepSite web server is available at http://pepsite2.russelllab.org.

DOI: 10.1093/nar/gks398

PMCID: PMC3394340

PMID: 22600738 [Indexed for MEDLINE]

1394. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W214-21. doi:

10.1093/nar/gks435. Epub 2012 May 16.

ProBiS-2012: web server and web services for detection of structurally similar

binding sites in proteins.

Konc J(1), Janezic D.

Author information:

(1)National Institute of Chemistry, Ljubljana, Slovenia.

The ProBiS web server is a web server for detection of structurally similar

binding sites in the PDB and for local pairwise alignment of protein structures.

In this article, we present a new version of the ProBiS web server that is 10

times faster than earlier versions, due to the efficient parallelization of the

ProBiS algorithm, which now allows significantly faster comparison of a protein

query against the PDB and reduces the calculation time for scanning the entire

PDB from hours to minutes. It also features new web services, and an improved

user interface. In addition, the new web server is united with the

ProBiS-Database and thus provides instant access to pre-calculated protein

similarity profiles for over 29 000 non-redundant protein structures. The ProBiS

web server is particularly adept at detection of secondary binding sites in

proteins. It is freely available at http://probis.cmm.ki.si/old-version, and the

new ProBiS web server is at http://probis.cmm.ki.si.

DOI: 10.1093/nar/gks435

PMCID: PMC3394329

PMID: 22600737 [Indexed for MEDLINE]

1395. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W376-9. doi: 10.1093/nar/gks437.

Epub 2012 May 16.

ASEB: a web server for KAT-specific acetylation site prediction.

Wang L(1), Du Y, Lu M, Li T.

Author information:

(1)Institute of Systems Biomedicine, School of Basic Medical Sciences, Peking

University Health Science Center, Beijing 100191, China.

Protein lysine acetylation plays an important role in the normal functioning of

cells, including gene expression regulation, protein stability and metabolism

regulation. Although large amounts of lysine acetylation sites have been

identified via large-scale mass spectrometry or traditional experimental methods,

the lysine (K)-acetyl-transferase (KAT) responsible for the acetylation of a

given protein or lysine site remains largely unknown due to the experimental

limitations of KAT substrate identification. Hence, the in silico prediction of

KAT-specific acetylation sites may provide direction for further experiments. In

our previous study, we developed the acetylation set enrichment based (ASEB)

computer program to predict which KAT-families are responsible for the

acetylation of a given protein or lysine site. In this article, we provide

KAT-specific acetylation site prediction as a web service. This web server not

only provides the online tool and R package for the method in our previous study,

but several useful services are also included, such as the integration of

protein-protein interaction information to enhance prediction accuracy. This web

server can be freely accessed at http://cmbi.bjmu.edu.cn/huac.

DOI: 10.1093/nar/gks437

PMCID: PMC3394258

PMID: 22600735 [Indexed for MEDLINE]

1396. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W29-34. doi: 10.1093/nar/gks412.

Epub 2012 May 16.

Rtips: fast and accurate tools for RNA 2D structure prediction using integer

programming.

Kato Y(1), Sato K, Asai K, Akutsu T.

Author information:

(1)Graduate School of Information Science, Nara Institute of Science and

Technology (NAIST), 8916-5 Takayama, Ikoma, Nara 630-0192, Japan.

ykato@is.naist.jp

We present a web-based tool set Rtips for fast and accurate prediction of RNA 2D

complex structures. Rtips comprises two computational tools based on integer

programming, IPknot for predicting RNA secondary structures with pseudoknots and

RactIP for predicting RNA-RNA interactions with kissing hairpins. Both servers

can run much faster than existing services with the same purpose on large data

sets as well as being at least comparable in prediction accuracy. The Rtips web

server along with the stand-alone programs is freely accessible at

http://rna.naist.jp/.

DOI: 10.1093/nar/gks412

PMCID: PMC3394313

PMID: 22600734 [Indexed for MEDLINE]

1397. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W288-93. doi:

10.1093/nar/gks419. Epub 2012 May 11.

PEP-FOLD: an updated de novo structure prediction server for both linear and

disulfide bonded cyclic peptides.

Thévenet P(1), Shen Y, Maupetit J, Guyon F, Derreumaux P, Tufféry P.

Author information:

(1)INSERM, U973, MTi, Paris, France.

In the context of the renewed interest of peptides as therapeutics, it is

important to have an on-line resource for 3D structure prediction of peptides

with well-defined structures in aqueous solution. We present an updated version

of PEP-FOLD allowing the treatment of both linear and disulphide bonded cyclic

peptides with 9-36 amino acids. The server makes possible to define disulphide

bonds and any residue-residue proximity under the guidance of the biologists.

Using a benchmark of 34 cyclic peptides with one, two and three disulphide bonds,

the best PEP-FOLD models deviate by an average RMS of 2.75 Å from the full NMR

structures. Using a benchmark of 37 linear peptides, PEP-FOLD locates

lowest-energy conformations deviating by 3 Å RMS from the NMR rigid cores. The

evolution of PEP-FOLD comes as a new on-line service to supersede the previous

server. The server is available at:

http://bioserv.rpbs.univ-paris-diderot.fr/PEP-FOLD.

DOI: 10.1093/nar/gks419

PMCID: PMC3394260

PMID: 22581768 [Indexed for MEDLINE]

1398. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W96-W103. doi:

10.1093/nar/gks422. Epub 2012 May 10.

QGRS-H Predictor: a web server for predicting homologous quadruplex forming

G-rich sequence motifs in nucleotide sequences.

Menendez C(1), Frees S, Bagga PS.

Author information:

(1)Bioinformatics and Computer Science, School of Theoretical and Applied

Science, Ramapo College of New Jersey, Mahwah, NJ 07430, USA.

Naturally occurring G-quadruplex structural motifs, formed by guanine-rich

nucleic acids, have been reported in telomeric, promoter and transcribed regions

of mammalian genomes. G-quadruplex structures have received significant attention

because of growing evidence for their role in important biological processes,

human disease and as therapeutic targets. Lately, there has been much interest in

the potential roles of RNA G-quadruplexes as cis-regulatory elements of

post-transcriptional gene expression. Large-scale computational genomics studies

on G-quadruplexes have difficulty validating their predictions without laborious

testing in 'wet' labs. We have developed a bioinformatics tool, QGRS-H Predictor

that can map and analyze conserved putative Quadruplex forming 'G'-Rich Sequences

(QGRS) in mRNAs, ncRNAs and other nucleotide sequences, e.g. promoter, telomeric

and gene flanking regions. Identifying conserved regulatory motifs helps validate

computations and enhances accuracy of predictions. The QGRS-H Predictor is

particularly useful for mapping homologous G-quadruplex forming sequences as

cis-regulatory elements in the context of 5'- and 3'-untranslated regions, and

CDS sections of aligned mRNA sequences. QGRS-H Predictor features highly

interactive graphic representation of the data. It is a unique and user-friendly

application that provides many options for defining and studying G-quadruplexes.

The QGRS-H Predictor can be freely accessed at:

http://quadruplex.ramapo.edu/qgrs/app/start.

DOI: 10.1093/nar/gks422

PMCID: PMC3394323

PMID: 22576365 [Indexed for MEDLINE]

1399. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W110-6. doi: 10.1093/nar/gks365.

Epub 2012 May 8.

LAHEDES: the LAGLIDADG homing endonuclease database and engineering server.

Taylor GK(1), Petrucci LH, Lambert AR, Baxter SK, Jarjour J, Stoddard BL.

Author information:

(1)Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle,

WA 98109, USA.

LAGLIDADG homing endonucleases (LHEs) are DNA cleaving enzymes, also termed

'meganucleases' that are employed as gene-targeting reagents. This use of LHEs

requires that their DNA specificity be altered to match sequences in genomic

targets. The choice of the most appropriate LHE to target a particular gene is

facilitated by the growing number of such enzymes with well-characterized

activities and structures. 'LAHEDES' (The LAGLIDADG Homing Endonuclease Database

and Engineering Server) provides both an online archive of LHEs with validated

DNA cleavage specificities and DNA-binding interactions, as well as a tool for

the identification of DNA sequences that might be targeted by various LHEs.

Searches can be performed using four separate scoring algorithms and user-defined

choices of LHE scaffolds. The webserver subsequently provides information

regarding clusters of amino acids that should be interrogated during engineering

and selection experiments. The webserver is fully open access and can be found at

http://homingendonuclease.net.

DOI: 10.1093/nar/gks365

PMCID: PMC3394308

PMID: 22570419 [Indexed for MEDLINE]

1400. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W537-41. doi:

10.1093/nar/gks375. Epub 2012 May 8.

H++ 3.0: automating pK prediction and the preparation of biomolecular structures

for atomistic molecular modeling and simulations.

Anandakrishnan R(1), Aguilar B, Onufriev AV.

Author information:

(1)Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA.

The accuracy of atomistic biomolecular modeling and simulation studies depend on

the accuracy of the input structures. Preparing these structures for an atomistic

modeling task, such as molecular dynamics (MD) simulation, can involve the use of

a variety of different tools for: correcting errors, adding missing atoms,

filling valences with hydrogens, predicting pK values for titratable amino acids,

assigning predefined partial charges and radii to all atoms, and generating force

field parameter/topology files for MD. Identifying, installing and effectively

using the appropriate tools for each of these tasks can be difficult for novice

and time-consuming for experienced users. H++ (http://biophysics.cs.vt.edu/) is a

free open-source web server that automates the above key steps in the preparation

of biomolecular structures for molecular modeling and simulations. H++ also

performs extensive error and consistency checking, providing error/warning

messages together with the suggested corrections. In addition to numerous minor

improvements, the latest version of H++ includes several new capabilities and

options: fix erroneous (flipped) side chain conformations for HIS, GLN and ASN,

include a ligand in the input structure, process nucleic acid structures and

generate a solvent box with specified number of common ions for explicit solvent

MD.

DOI: 10.1093/nar/gks375

PMCID: PMC3394296

PMID: 22570416 [Indexed for MEDLINE]

1401. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W542-6. doi: 10.1093/nar/gks373.

Epub 2012 May 8.

RPF: a quality assessment tool for protein NMR structures.

Huang YJ(1), Rosato A, Singh G, Montelione GT.

Author information:

(1)Center for Advanced Biotechnology and Medicine, Northeast Structural Genomic

Consortium, Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854, USA.

We describe the RPF web server, a quality assessment tool for protein NMR

structures. The RPF server measures the 'goodness-of-fit' of the 3D structure

with NMR chemical shift and unassigned NOESY data, and calculates a

discrimination power (DP) score, which estimates the differences between the fits

of the query structures and random coil structures to these experimental data.

The DP-score is an accuracy predictor of the query structure. The RPF server also

maps local structure quality measures onto the 3D structure using an online

molecular viewer, and onto the NMR spectra, allowing refinement of the structure

and/or NOESY peak list data. The RPF server is available at:

http://nmr.cabm.rutgers.edu/rpf.

DOI: 10.1093/nar/gks373

PMCID: PMC3394279

PMID: 22570414 [Indexed for MEDLINE]

1402. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W123-6. doi: 10.1093/nar/gks386.

Epub 2012 May 8.

SurvNet: a web server for identifying network-based biomarkers that most

correlate with patient survival data.

Li J(1), Roebuck P, Grünewald S, Liang H.

Author information:

(1)Chinese Academy of Sciences Key Laboratory of Computational Biology, Chinese

Academy of Sciences and Max Planck Society (CAS-MPG) Partner Institute for

Computational Biology, Shanghai Institutes for Biological Sciences, Chinese

Academy of Sciences, Shanghai 200031, P.R. China.

An important task in biomedical research is identifying biomarkers that correlate

with patient clinical data, and these biomarkers then provide a critical

foundation for the diagnosis and treatment of disease. Conventionally, such an

analysis is based on individual genes, but the results are often noisy and

difficult to interpret. Using a biological network as the searching platform,

network-based biomarkers are expected to be more robust and provide deep insights

into the molecular mechanisms of disease. We have developed a novel

bioinformatics web server for identifying network-based biomarkers that most

correlate with patient survival data, SurvNet. The web server takes three input

files: one biological network file, representing a gene regulatory or protein

interaction network; one molecular profiling file, containing any type of gene-

or protein-centred high-throughput biological data (e.g. microarray expression

data or DNA methylation data); and one patient survival data file (e.g. patients'

progression-free survival data). Given user-defined parameters, SurvNet will

automatically search for subnetworks that most correlate with the observed

patient survival data. As the output, SurvNet will generate a list of network

biomarkers and display them through a user-friendly interface. SurvNet can be

accessed at http://bioinformatics.mdanderson.org/main/SurvNet.

DOI: 10.1093/nar/gks386

PMCID: PMC3394266

PMID: 22570412 [Indexed for MEDLINE]

1403. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W209-13. doi:

10.1093/nar/gks396. Epub 2012 May 8.

RF-Cloning.org: an online tool for the design of restriction-free cloning

projects.

Bond SR(1), Naus CC.

Author information:

(1)Department of Cellular and Physiological Science, Life Sciences Institute,

University of British Columbia, Vancouver, British Columbia, V6T 1Z3, Canada.

biologyguy@gmail.com

Restriction-free cloning (RF-cloning) is a PCR-based technology that expands on

the QuikChange™ mutagenesis process originally popularized by Stratagene in the

mid-1990s, and allows the insertion of essentially any sequence into any plasmid

at any location. While RF-cloning is a powerful tool for the design of custom

plasmids when restriction sites are not conveniently situated, manually designing

the requisite primers can be tedious and error prone. We present here a

web-service that automates the primer design process, along with a user interface

that includes a number of useful tools for managing both the input sequences and

the resulting outputs. RF-Cloning is free and open to all users, and can be

accessed at http://www.rf-cloning.org.

DOI: 10.1093/nar/gks396

PMCID: PMC3394257

PMID: 22570410 [Indexed for MEDLINE]

1404. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W458-65. doi:

10.1093/nar/gks380. Epub 2012 May 8.

GPSy: a cross-species gene prioritization system for conserved biological

processes--application in male gamete development.

Britto R(1), Sallou O, Collin O, Michaux G, Primig M, Chalmel F.

Author information:

(1)Inserm Unité 1085-Irset, France.

We present gene prioritization system (GPSy), a cross-species gene prioritization

system that facilitates the arduous but critical task of prioritizing genes for

follow-up functional analyses. GPSy's modular design with regard to species, data

sets and scoring strategies enables users to formulate queries in a highly

flexible manner. Currently, the system encompasses 20 topics related to conserved

biological processes including male gamete development discussed in this article.

The web server-based tool is freely available at http://gpsy.genouest.org.

DOI: 10.1093/nar/gks380

PMCID: PMC3394256

PMID: 22570409 [Indexed for MEDLINE]

1405. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W323-8. doi: 10.1093/nar/gks376.

Epub 2012 May 7.

KoBaMIN: a knowledge-based minimization web server for protein structure

refinement.

Rodrigues JP(1), Levitt M, Chopra G.

Author information:

(1)Department of Structural Biology, 299 Campus Dr W, Fairchild Bldg, Room D100,

Stanford University, Stanford, CA 94305, USA.

The KoBaMIN web server provides an online interface to a simple, consistent and

computationally efficient protein structure refinement protocol based on

minimization of a knowledge-based potential of mean force. The server can be used

to refine either a single protein structure or an ensemble of proteins starting

from their unrefined coordinates in PDB format. The refinement method is

particularly fast and accurate due to the underlying knowledge-based potential

derived from structures deposited in the PDB; as such, the energy function

implicitly includes the effects of solvent and the crystal environment. Our

server allows for an optional but recommended step that optimizes stereochemistry

using the MESHI software. The KoBaMIN server also allows comparison of the

refined structures with a provided reference structure to assess the changes

brought about by the refinement protocol. The performance of KoBaMIN has been

benchmarked widely on a large set of decoys, all models generated at the seventh

worldwide experiments on critical assessment of techniques for protein structure

prediction (CASP7) and it was also shown to produce top-ranking predictions in

the refinement category at both CASP8 and CASP9, yielding consistently good

results across a broad range of model quality values. The web server is fully

functional and freely available at http://csb.stanford.edu/kobamin.

DOI: 10.1093/nar/gks376

PMCID: PMC3394243

PMID: 22564897 [Indexed for MEDLINE]

1406. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W127-33. doi:

10.1093/nar/gks374. Epub 2012 May 2.

MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis.

Xia J(1), Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS.

Author information:

(1)Department of Biological Sciences, National Institute for Nanotechnology

(NINT), Edmonton, AB, Canada T6G 2E8.

First released in 2009, MetaboAnalyst (www.metaboanalyst.ca) was a relatively

simple web server designed to facilitate metabolomic data processing and

statistical analysis. With continuing advances in metabolomics along with

constant user feedback, it became clear that a substantial upgrade to the

original server was necessary. MetaboAnalyst 2.0, which is the successor to

MetaboAnalyst, represents just such an upgrade. MetaboAnalyst 2.0 now contains

dozens of new features and functions including new procedures for data filtering,

data editing and data normalization. It also supports multi-group data analysis,

two-factor analysis as well as time-series data analysis. These new functions

have also been supplemented with: (i) a quality-control module that allows users

to evaluate their data quality before conducting any analysis, (ii) a functional

enrichment analysis module that allows users to identify biologically meaningful

patterns using metabolite set enrichment analysis and (iii) a metabolic pathway

analysis module that allows users to perform pathway analysis and visualization

for 15 different model organisms. In developing MetaboAnalyst 2.0 we have also

substantially improved its graphical presentation tools. All images are now

generated using anti-aliasing and are available over a range of resolutions,

sizes and formats (PNG, TIFF, PDF, PostScript, or SVG). To improve its

performance, MetaboAnalyst 2.0 is now hosted on a much more powerful server with

substantially modified code to take advantage the server's multi-core CPUs for

computationally intensive tasks. MetaboAnalyst 2.0 also maintains a collection of

50 or more FAQs and more than a dozen tutorials compiled from user queries and

requests. A downloadable version of MetaboAnalyst 2.0, along detailed

instructions for local installation is now available as well.

DOI: 10.1093/nar/gks374

PMCID: PMC3394314

PMID: 22553367 [Indexed for MEDLINE]

1407. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W303-9. doi: 10.1093/nar/gks362.

Epub 2012 May 2.

CSA: comprehensive comparison of pairwise protein structure alignments.

Wohlers I(1), Malod-Dognin N, Andonov R, Klau GW.

Author information:

(1)Life Sciences Group, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG

Amsterdam, The Netherlands. I.Wohlers@cwi.nl

CSA is a web server for the computation, evaluation and comprehensive comparison

of pairwise protein structure alignments. Its exact alignment engine computes

either optimal, top-scoring alignments or heuristic alignments with quality

guarantee for the inter-residue distance-based scorings of contact map overlap,

PAUL, DALI and MATRAS. These and additional, uploaded alignments are compared

using a number of quality measures and intuitive visualizations. CSA brings new

insight into the structural relationship of the protein pairs under investigation

and is a valuable tool for studying structural similarities. It is available at

http://csa.project.cwi.nl.

DOI: 10.1093/nar/gks362

PMCID: PMC3394275

PMID: 22553365 [Indexed for MEDLINE]

1408. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W298-302. doi:

10.1093/nar/gks361. Epub 2012 May 2.

DSP: a protein shape string and its profile prediction server.

Sun J(1), Tang S, Xiong W, Cong P, Li T.

Author information:

(1)Department of Chemistry, Tongji University, 1239 Siping Road, Shanghai 200092,

China.

Many studies have demonstrated that shape string is an extremely important

structure representation, since it is more complete than the classical secondary

structure. The shape string provides detailed information also in the regions

denoted random coil. But few services are provided for systematic analysis of

protein shape string. To fill this gap, we have developed an accurate shape

string predictor based on two innovative technologies: a knowledge-driven

sequence alignment and a sequence shape string profile method. The performance on

blind test data demonstrates that the proposed method can be used for accurate

prediction of protein shape string. The DSP server provides both predicted shape

string and sequence shape string profile for each query sequence. Using this

information, the users can compare protein structure or display protein evolution

in shape string space. The DSP server is available at both

http://cheminfo.tongji.edu.cn/dsp/ and its main mirror

http://chemcenter.tongji.edu.cn/dsp/.

DOI: 10.1093/nar/gks361

PMCID: PMC3394270

PMID: 22553364 [Indexed for MEDLINE]

1409. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W65-70. doi: 10.1093/nar/gks364.

Epub 2012 Apr 28.

SNPnexus: a web server for functional annotation of novel and publicly known

genetic variants (2012 update).

Dayem Ullah AZ(1), Lemoine NR, Chelala C.

Author information:

(1)Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, UK.

Broader functional annotation of single nucleotide variations is a valuable mean

for prioritizing targets in further disease studies and large-scale genotyping

projects. We originally developed SNPnexus to assess the potential significance

of known and novel SNPs on the major transcriptome, proteome, regulatory and

structural variation models in order to identify the phenotypically important

variants. Being committed to providing continuous support to the scientific

community, we have substantially improved SNPnexus over time by incorporating a

broader range of variations such as insertions/deletions, block substitutions,

IUPAC codes submission and region-based analysis, expanding the query size limit,

and most importantly including additional categories for the assessment of

functional impact. SNPnexus provides a comprehensive set of annotations for

genomic variation data by characterizing related functional consequences at the

transcriptome/proteome levels of seven major annotation systems with in-depth

analysis of potential deleterious effects, inferring physical and cytogenetic

mapping, reporting information on HapMap genotype/allele data, finding overlaps

with potential regulatory elements, structural variations and conserved elements,

and retrieving links with previously reported genetic disease studies. SNPnexus

has a user-friendly web interface with an improved query structure, enhanced

functional annotation categories and flexible output presentation making it

practically useful for biologists. SNPnexus is freely available at

http://www.snp-nexus.org.

DOI: 10.1093/nar/gks364

PMCID: PMC3394262

PMID: 22544707 [Indexed for MEDLINE]

1410. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W59-64. doi: 10.1093/nar/gks310.

Epub 2012 Apr 6.

SAVoR: a server for sequencing annotation and visualization of RNA structures.

Li F(1), Ryvkin P, Childress DM, Valladares O, Gregory BD, Wang LS.

Author information:

(1)Genomics and Computational Biology Graduate Group, Perelman School of

Medicine, University of Pennsylvania, PA 19104, USA.

RNA secondary structure is required for the proper regulation of the cellular

transcriptome. This is because the functionality, processing, localization and

stability of RNAs are all dependent on the folding of these molecules into

intricate structures through specific base pairing interactions encoded in their

primary nucleotide sequences. Thus, as the number of RNA sequencing (RNA-seq)

data sets and the variety of protocols for this technology grow rapidly, it is

becoming increasingly pertinent to develop tools that can analyze and visualize

this sequence data in the context of RNA secondary structure. Here, we present

Sequencing Annotation and Visualization of RNA structures (SAVoR), a web server,

which seamlessly links RNA structure predictions with sequencing data and genomic

annotations to produce highly informative and annotated models of RNA secondary

structure. SAVoR accepts read alignment data from RNA-seq experiments and

computes a series of per-base values such as read abundance and sequence variant

frequency. These values can then be visualized on a customizable secondary

structure model. SAVoR is freely available at http://tesla.pcbi.upenn.edu/savor.

DOI: 10.1093/nar/gks310

PMCID: PMC3394343

PMID: 22492627 [Indexed for MEDLINE]

1411. Proteins. 2012 Jul;80(7):1791-7. doi: 10.1002/prot.24074. Epub 2012 May 8.

Sann: solvent accessibility prediction of proteins by nearest neighbor method.

Joo K(1), Lee SJ, Lee J.

Author information:

(1)Center for In Silico Protein Science, Korea Institute for Advanced Study,

Korea.

We present a method to predict the solvent accessibility of proteins which is

based on a nearest neighbor method applied to the sequence profiles. Using the

method, continuous real-value prediction as well as two-state and three-state

discrete predictions can be obtained. The method utilizes the z-score value of

the distance measure in the feature vector space to estimate the relative

contribution among the k-nearest neighbors for prediction of the discrete and

continuous solvent accessibility. The Solvent accessibility database is

constructed from 5717 proteins extracted from PISCES culling server with the

cutoff of 25% sequence identities. Using optimal parameters, the prediction

accuracies (for discrete predictions) of 78.38% (two-state prediction with the

threshold of 25%), 65.1% (three-state prediction with the thresholds of 9 and

36%), and the Pearson correlation coefficient (between the predicted and true

RSA's for continuous prediction) of 0.676 are achieved An independent benchmark

test was performed with the CASP8 targets where we find that the proposed method

outperforms existing methods. The prediction accuracies are 80.89% (for two state

prediction with the threshold of 25%), 67.58% (three-state prediction), and the

Pearson correlation coefficient of 0.727 (for continuous prediction) with mean

absolute error of 0.148. We have also investigated the effect of increasing

database sizes on the prediction accuracy, where additional improvement in the

accuracy is observed as the database size increases. The SANN web server is

available at http://lee.kias.re.kr/~newton/sann/.

Copyright © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/prot.24074

PMID: 22434533 [Indexed for MEDLINE]

1412. J Chem Inf Model. 2012 Jun 25;52(6):1674-85. doi: 10.1021/ci300123x. Epub 2012

Jun 13.

3-D QSAutogrid/R: an alternative procedure to build 3-D QSAR models.

Methodologies and applications.

Ballante F(1), Ragno R.

Author information:

(1)Rome Center for Molecular Design, Dipartimento di Chimica e Tecnologie del

Farmaco, Sapienza Università di Roma, P. le A. Moro 5, 00185, Rome, Italy.

Since it first appeared in 1988 3-D QSAR has proved its potential in the field of

drug design and activity prediction. Although thousands of citations now exist in

3-D QSAR, its development was rather slow with the majority of new 3-D QSAR

applications just extensions of CoMFA. An alternative way to build 3-D QSAR

models, based on an evolution of software, has been named 3-D QSAutogrid/R and

has been developed to use only software freely available to academics. 3-D

QSAutogrid/R covers all the main features of CoMFA and GRID/GOLPE with

implementation by multiprobe/multiregion variable selection (MPGRS) that improves

the simplification of interpretation of the 3-D QSAR map. The methodology is

based on the integration of the molecular interaction fields as calculated by

AutoGrid and the R statistical environment that can be easily coupled with many

free graphical molecular interfaces such as UCSF-Chimera, AutoDock Tools, JMol,

and others. The description of each R package is reported in detail, and, to

assess its validity, 3-D QSAutogrid/R has been applied to three molecular data

sets of which either CoMFA or GRID/GOLPE models were reported in order to compare

the results. 3-D QSAutogrid/R has been used as the core engine to prepare more

that 240 3-D QSAR models forming the very first 3-D QSAR server ( www.3d-qsar.com

) with its code freely available through R-Cran distribution.

DOI: 10.1021/ci300123x

PMID: 22643034 [Indexed for MEDLINE]

1413. Anal Biochem. 2012 Jun 15;425(2):117-9. doi: 10.1016/j.ab.2012.03.015. Epub 2012

Mar 27.

PseAAC-Builder: a cross-platform stand-alone program for generating various

special Chou's pseudo-amino acid compositions.

Du P(1), Wang X, Xu C, Gao Y.

Author information:

(1)School of Computer Science and Technology, Tianjin University, Tianjin 300072,

China. pufengdu@gmail.com

The pseudo-amino acid composition has been widely used to convert complicated

protein sequences with various lengths to fixed length digital feature vectors

while keeping considerable sequence order information. However, so far the only

software available to the public is the web server PseAAC

(http://www.csbio.sjtu.edu.cn/bioinf/PseAAC), which has some limitations in

dealing with large-scale datasets. Here, we propose a new cross-platform

stand-alone software program, called PseAAC-Builder (http://www.pseb.sf.net),

which can be used to generate various modes of Chou's pseudo-amino acid

composition in a much more efficient and flexible way. It is anticipated that

PseAAC-Builder may become a useful tool for studying various protein attributes.

Copyright © 2012 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ab.2012.03.015

PMID: 22459120 [Indexed for MEDLINE]

1414. Bioinformatics. 2012 Jun 15;28(12):i90-6. doi: 10.1093/bioinformatics/bts233.

Ranking models of transmembrane β-barrel proteins using Z-coordinate predictions.

Hayat S(1), Elofsson A.

Author information:

(1)Center for Biomembrane Research, Department of Biochemistry and Biophysics,

Stockholm Bioinformatics Center, Science for Life Laboratory, Swedish E-science

Research Center, Stockholm University, SE-10691 Stockholm, Sweden.

MOTIVATION: Transmembrane β-barrels exist in the outer membrane of gram-negative

bacteria as well as in chloroplast and mitochondria. They are often involved in

transport processes and are promising antimicrobial drug targets. Structures of

only a few β-barrel protein families are known. Therefore, a method that could

automatically generate such models would be valuable. The symmetrical arrangement

of the barrels suggests that an approach based on idealized geometries may be

successful.

RESULTS: Here, we present tobmodel; a method for generating 3D models of β-barrel

transmembrane proteins. First, alternative topologies are obtained from the

BOCTOPUS topology predictor. Thereafter, several 3D models are constructed by

using different angles of the β-sheets. Finally, the best model is selected based

on agreement with a novel predictor, ZPRED3, which predicts the distance from the

center of the membrane for each residue, i.e. the Z-coordinate. The Z-coordinate

prediction has an average error of 1.61 Å. Tobmodel predicts the correct topology

for 75% of the proteins in the dataset which is a slight improvement over

BOCTOPUS alone. More importantly, however, tobmodel provides a Cα template with

an average RMSD of 7.24 Å from the native structure.

AVAILABILITY: Tobmodel is freely available as a web server at:

http://tobmodel.cbr.su.se/. The datasets used for training and evaluations are

also available from this site.

DOI: 10.1093/bioinformatics/bts233

PMCID: PMC3371865

PMID: 22689784 [Indexed for MEDLINE]

1415. Bioinformatics. 2012 Jun 15;28(12):1562-70. doi: 10.1093/bioinformatics/bts195.

Epub 2012 Apr 23.

MMFPh: a maximal motif finder for phosphoproteomics datasets.

Wang T(1), Kettenbach AN, Gerber SA, Bailey-Kellogg C.

Author information:

(1)Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA.

Comment in

Bioinformatics. 2012 Aug 15;28(16):2211-2; author reply 2213.

MOTIVATION: Protein phosphorylation, driven by specific recognition of substrates

by kinases and phosphatases, plays central roles in a variety of important

cellular processes such as signaling and enzyme activation. Mass spectrometry

enables the determination of phosphorylated peptides (and thereby proteins) in

scenarios ranging from targeted in vitro studies to in vivo cell lysates under

particular conditions. The characterization of commonalities among identified

phosphopeptides provides insights into the specificities of the kinases involved

in a study. Several algorithms have been developed to uncover linear motifs

representing position-specific amino acid patterns in sets of phosphopeptides. To

more fully capture the available information, reduce sensitivity to both

parameter choices and natural experimental variation, and develop more precise

characterizations of kinase specificities, it is necessary to determine all

statistically significant motifs represented in a dataset.

RESULTS: We have developed MMFPh (Maximal Motif Finder for Phosphoproteomics

datasets), which extends the approach of the popular phosphorylation motif

software Motif-X (Schwartz and Gygi, 2005) to identify all statistically

significant motifs and return the maximal ones (those not subsumed by motifs with

more fixed amino acids). In tests with both synthetic and experimental data, we

show that MMFPh finds important motifs missed by the greedy approach of Motif-X,

while also finding more motifs that are more characteristic of the dataset

relative to the background proteome. Thus MMFPh is in some sense both more

sensitive and more specific in characterizing the involved kinases. We also show

that MMFPh compares favorably to other recent methods for finding phosphorylation

motifs. Furthermore, MMFPh is less dependent on parameter choices. We support

this powerful new approach with a web interface so that it may become a useful

tool for studies of kinase specificity and phosphorylation site prediction.

AVAILABILITY: A web server is at www.cs.dartmouth.edu/~cbk/.

DOI: 10.1093/bioinformatics/bts195

PMCID: PMC3371830

PMID: 22531218 [Indexed for MEDLINE]

1416. Bioinformatics. 2012 Jun 15;28(12):1655-7. doi: 10.1093/bioinformatics/bts200.

Epub 2012 Apr 23.

DelPhi web server v2: incorporating atomic-style geometrical figures into the

computational protocol.

Smith N(1), Witham S, Sarkar S, Zhang J, Li L, Li C, Alexov E.

Author information:

(1)Computational Biophysics and Bioinformatics, Department of Physics, Clemson

University, Clemson, SC 29634, USA. nsmith@clemson.edu

A new edition of the DelPhi web server, DelPhi web server v2, is released to

include atomic presentation of geometrical figures. These geometrical objects can

be used to model nano-size objects together with real biological macromolecules.

The position and size of the object can be manipulated by the user in real time

until desired results are achieved. The server fixes structural defects, adds

hydrogen atoms and calculates electrostatic energies and the corresponding

electrostatic potential and ionic distributions.AVAILABILITY AND IMPLEMENTATION:

The web server follows a client-server architecture built on PHP and HTML and

utilizes DelPhi software. The computation is carried out on supercomputer cluster

and results are given back to the user via http protocol, including the ability

to visualize the structure and corresponding electrostatic potential via Jmol

implementation. The DelPhi web server is available from

http://compbio.clemson.edu/delphi\_webserver.

DOI: 10.1093/bioinformatics/bts200

PMCID: PMC3371833

PMID: 22531215 [Indexed for MEDLINE]

1417. Bioinformatics. 2012 Jun 15;28(12):1643-4. doi: 10.1093/bioinformatics/bts201.

Epub 2012 Apr 23.

NuST: analysis of the interplay between nucleoid organization and gene

expression.

Scolari VF(1), Zarei M, Osella M, Lagomarsino MC.

Author information:

(1)Genomic Physics Group, UMR 7238 CNRS Génomique des Microorganismes, Université

Pierre et Marie Curie, 15 rue de L'École de Médecine, 75006, Paris, France.

vittore.scolari@upmc.fr

Different experimental results suggest the presence of an interplay between

global transcriptional regulation and chromosome spatial organization in

bacteria. The identification and clear visualization of spatial clusters of

contiguous genes targeted by specific DNA-binding proteins or sensitive to

nucleoid perturbations can elucidate links between nucleoid structure and gene

expression patterns. Similarly, statistical analysis to assess correlations

between results from independent experiments can provide the integrated analysis

needed in this line of research. NuST (Nucleoid Survey tools), based on the

Escherichia coli genome, gives the non-expert the possibility to analyze the

aggregation of genes or loci sets along the genome coordinate, at different

scales of observation. It is useful to discover correlations between different

sources of data (e.g. expression, binding or genomic data) and genome

organization. A user can use it on datasets in the form of gene lists coming from

his/her own experiments or bioinformatic analyses, but also make use of the

internal database, which collects data from many published studies.AVAILABILITY

AND IMPLEMENTATION: NuST is a web server (available at

http://www.lgm.upmc.fr/nust/). The website is implemented in PHP, SQLite and

Ajax, with all major browsers supported, while the core algorithms are optimized

and implemented in C. NuST has an extensive help page and provides a direct

visualization of results as well as different downloadable file formats. A

template Perl code for automated access to the web server can be downloaded at

http://www.lgm.upmc.fr/nust/downloads/, in order to allow the users to use NuST

in systematic bioinformatic analyses.

DOI: 10.1093/bioinformatics/bts201

PMID: 22531214 [Indexed for MEDLINE]

1418. Bioinformatics. 2012 Jun 15;28(12):1579-85. doi: 10.1093/bioinformatics/bts182.

Epub 2012 Apr 11.

Ligand-binding site prediction using ligand-interacting and binding site-enriched

protein triangles.

Xie ZR(1), Hwang MJ.

Author information:

(1)Institute of Biomedical Informatics, National Yang-Ming University, Taipei

112, Taiwan.

MOTIVATION: Knowledge about the site at which a ligand binds provides an

important clue for predicting the function of a protein and is also often a

prerequisite for performing docking computations in virtual drug design and

screening. We have previously shown that certain ligand-interacting triangles of

protein atoms, called protein triangles, tend to occur more frequently at

ligand-binding sites than at other parts of the protein.

RESULTS: In this work, we describe a new ligand-binding site prediction method

that was developed based on binding site-enriched protein triangles. The new

method was tested on 2 benchmark datasets and on 19 targets from two recent

community-based studies of such predictions, and excellent results were obtained.

Where comparisons were made, the success rates for the new method for the first

predicted site were significantly better than methods that are not a

meta-predictor. Further examination showed that, for most of the unsuccessful

predictions, the pocket of the ligand-binding site was identified, but not the

site itself, whereas for some others, the failure was not due to the method

itself but due to the use of an incorrect biological unit in the structure

examined, although using correct biological units would not necessarily improve

the prediction success rates. These results suggest that the new method is a

valuable new addition to a suite of existing structure-based bioinformatics tools

for studies of molecular recognition and related functions of proteins in

post-genomics research.

AVAILABILITY: The executable binaries and a web server for our method are

available from http://sourceforge.net/projects/msdock/ and

http://lise.ibms.sinica.edu.tw, respectively, free for academic users.

DOI: 10.1093/bioinformatics/bts182

PMID: 22495747 [Indexed for MEDLINE]

1419. BMC Bioinformatics. 2012 Jun 7;13:123.

Quartet decomposition server: a platform for analyzing phylogenetic trees.

Mao F(1), Williams D, Zhaxybayeva O, Poptsova M, Lapierre P, Gogarten JP, Xu Y.

Author information:

(1)Department of Biochemistry and Molecular Biology, University of Georgia, 120

Green St, Athens, GA 30622, USA.

BACKGROUND: The frequent exchange of genetic material among prokaryotes means

that extracting a majority or plurality phylogenetic signal from many gene

families, and the identification of gene families that are in significant

conflict with the plurality signal is a frequent task in comparative genomics,

and especially in phylogenomic analyses. Decomposition of gene trees into

embedded quartets (unrooted trees each with four taxa) is a convenient and

statistically powerful technique to address this challenging problem. This

approach was shown to be useful in several studies of completely sequenced

microbial genomes.

RESULTS: We present here a web server that takes a collection of gene

phylogenies, decomposes them into quartets, generates a Quartet Spectrum, and

draws a split network. Users are also provided with various data download options

for further analyses. Each gene phylogeny is to be represented by an assessment

of phylogenetic information content, such as sets of trees reconstructed from

bootstrap replicates or sampled from a posterior distribution. The Quartet

Decomposition server is accessible at http://quartets.uga.edu.

CONCLUSIONS: The Quartet Decomposition server presented here provides a

convenient means to perform Quartet Decomposition analyses and will empower users

to find statistically supported phylogenetic conflicts.

DOI: 10.1186/1471-2105-13-123

PMCID: PMC3447714

PMID: 22676320 [Indexed for MEDLINE]

1420. Amino Acids. 2012 Jun;42(6):2363-71. doi: 10.1007/s00726-011-0978-z. Epub 2011

Jul 24.

Prediction of protein-protein interactions between Ralstonia solanacearum and

Arabidopsis thaliana.

Li ZG(1), He F, Zhang Z, Peng YL.

Author information:

(1)State Key Laboratory of Agrobiotechnology, China Agricultural University,

Beijing, 100193, China.

Ralstonia solanacearum is a devastating bacterial pathogen that has an unusually

wide host range. R. solanacearum, together with Arabidopsis thaliana, has become

a model system for studying the molecular basis of plant-pathogen interactions.

Protein-protein interactions (PPIs) play a critical role in the infection

process, and some PPIs can initiate a plant defense response. However,

experimental investigations have rarely addressed such PPIs. Using two

computational methods, the interolog and the domain-based methods, we predicted

3,074 potential PPIs between 119 R. solanacearum and 1,442 A. thaliana proteins.

Interestingly, we found that the potential pathogen-targeted proteins are more

important in the A. thaliana PPI network. To facilitate further studies, all

predicted PPI data were compiled into a database server called PPIRA

(http://protein.cau.edu.cn/ppira/). We hope that our work will provide new

insights for future research addressing the pathogenesis of R. solanacearum.

DOI: 10.1007/s00726-011-0978-z

PMID: 21786137 [Indexed for MEDLINE]

1421. Amino Acids. 2012 Jun;42(6):2243-9. doi: 10.1007/s00726-011-0964-5. Epub 2011 Jun

23.

Accurate prediction of protein structural class using auto covariance

transformation of PSI-BLAST profiles.

Liu T(1), Geng X, Zheng X, Li R, Wang J.

Author information:

(1)College of Information Sciences and Engineering, Shandong Agricultural

University, Taian, 271018, China.

Computational prediction of protein structural class based solely on sequence

data remains a challenging problem in protein science. Existing methods differ in

the protein sequence representation models and prediction engines adopted. In

this study, a powerful feature extraction method, which combines

position-specific score matrix (PSSM) with auto covariance (AC) transformation,

is introduced. Thus, a sample protein is represented by a series of discrete

components, which could partially incorporate the long-range sequence order

information and evolutionary information reflected from the PSI-BLAST profile. To

verify the performance of our method, jackknife cross-validation tests are

performed on four widely used benchmark datasets. Comparison of our results with

existing methods shows that our method provides the state-of-the-art performance

for structural class prediction. A Web server that implements the proposed method

is freely available at http://202.194.133.5/xinxi/AAC\_PSSM\_AC/index.htm.

DOI: 10.1007/s00726-011-0964-5

PMID: 21698456 [Indexed for MEDLINE]

1422. Bioinformatics. 2012 Jun 1;28(11):1538-9. doi: 10.1093/bioinformatics/bts179.

Epub 2012 Apr 11.

CheShift-2: graphic validation of protein structures.

Martin OA(1), Vila JA, Scheraga HA.

Author information:

(1)Universidad Nacional de San Luis, IMASL-CONICET, Ejército de Los Andes, San

Luis, Argentina.

The differences between observed and predicted (13)C(α) chemical shifts can be

used as a sensitive probe with which to detect possible local flaws in protein

structures. For this reason, we previously introduced CheShift, a Web server for

protein structure validation. Now, we present CheShift-2 in which a graphical

user interface is implemented to render such local flaws easily visible. A series

of applications to 15 ensembles of conformations illustrate the ability of

CheShift-2 to locate the main structural flaws rapidly and accurately on a

per-residue basis. Since accuracy plays a central role in CheShift predictions,

the treatment of histidine (His) is investigated here by exploring which form of

His should be used in CheShift-2.AVAILABILITY: CheShift-2 is free of charge for

academic use and can be accessed from www.cheshift.com

DOI: 10.1093/bioinformatics/bts179

PMCID: PMC3356844

PMID: 22495749 [Indexed for MEDLINE]

1423. Bioinformatics. 2012 Jun 1;28(11):1544-5. doi: 10.1093/bioinformatics/bts169.

Epub 2012 Apr 6.

PaGeFinder: quantitative identification of spatiotemporal pattern genes.

Pan JB(1), Hu SC, Wang H, Zou Q, Ji ZL.

Author information:

(1)Department of Chemical Biology, College of Chemistry and Chemical Engineering,

The Key Laboratory for Chemical Biology of Fujian Province, School of Life

Sciences, Xiamen University, Xiamen, Fujian, P R China.

Pattern Gene Finder (PaGeFinder) is a web-based server for on-line detection of

gene expression patterns from serial transcriptomic data generated by

high-throughput technologies like microarray or next-generation sequencing. Three

particular parameters, the specificity measure, the dispersion measure and the

contribution measure, were introduced and implemented in PaGeFinder to help

quantitative and interactive identification of pattern genes like housekeeping

genes, specific (selective) genes and repressed genes. Besides the on-line

computation service, the PaGeFinder also provides downloadable Java programs for

local detection of gene expression patterns.AVAILABILITY:

http://bioinf.xmu.edu.cn:8080/PaGeFinder/index.jsp

DOI: 10.1093/bioinformatics/bts169

PMCID: PMC3356841

PMID: 22492640 [Indexed for MEDLINE]

1424. FEBS J. 2012 Jun;279(12):2192-200. doi: 10.1111/j.1742-4658.2012.08603.x. Epub

2012 May 21.

PROSO II--a new method for protein solubility prediction.

Smialowski P(1), Doose G, Torkler P, Kaufmann S, Frishman D.

Author information:

(1)Department of Genome Oriented Bioinformatics, Technische Universität Muenchen,

Freising, Germany. pawelsm@gmail.com

Many fields of science and industry depend on efficient production of active

protein using heterologous expression in Escherichia coli. The solubility of

proteins upon expression is dependent on their amino acid sequence. Prediction of

solubility from sequence is therefore highly valuable. We present a novel

machine-learning-based model called PROSO II which makes use of new

classification methods and growth in experimental data to improve coverage and

accuracy of solubility predictions. The classification algorithm is organized as

a two-layered structure in which the output of a primary Parzen window model for

sequence similarity and a logistic regression classifier of amino acid k-mer

composition serve as input for a second-level logistic regression classifier.

Compared with previously published research our model is trained on five times

more data than used by any other method before (82 000 proteins). When tested on

a separate holdout set not used at any point of method development our server

attained the best results in comparison with other currently available methods:

accuracy 75.4%, Matthew's correlation coefficient 0.39, sensitivity 0.731,

specificity 0.759, gain (soluble) 2.263. In summary, due to utilization of

cutting edge machine learning technologies combined with the largest currently

available experimental data set the PROSO II server constitutes a substantial

improvement in protein solubility predictions. PROSO II is available at

http://mips.helmholtz-muenchen.de/prosoII.

© 2012 The Authors Journal compilation © 2012 FEBS.

DOI: 10.1111/j.1742-4658.2012.08603.x

PMID: 22536855 [Indexed for MEDLINE]

1425. Genomics. 2012 Jun;99(6):370-5. doi: 10.1016/j.ygeno.2012.04.002. Epub 2012 Apr

21.

Identification of mirtrons in rice using MirtronPred: a tool for predicting plant

mirtrons.

Joshi PK(1), Gupta D, Nandal UK, Khan Y, Mukherjee SK, Sanan-Mishra N.

Author information:

(1)Plant Molecular Biology Group, International Center for Genetic Engineering

and Biotechnology, Aruna Asaf Ali Marg, New Delhi, India.

Studies from flies and insects have reported the existence of a special class of

miRNA, called mirtrons that are produced from spliced-out introns in a

DROSHA-independent manner. The spliced-out lariat is debranched and refolded into

a stem-loop structure resembling the pre-miRNA, which can then be processed by

DICER into mature ~21 nt species. The mirtrons have not been reported from

plants. In this study, we present MirtronPred, a web based server to predict

mirtrons from intronic sequences. We have used the server to predict 70 mirtrons

in rice introns that were put through a stringent selection filter to shortlist

16 best sequences. The prediction accuracy was subsequently validated by northern

analysis and RT-PCR of a predicted Os-mirtron-109. The target sequences for this

mirtron were also found in the rice degradome database. The possible role of the

mirtron in rice regulon is discussed. The MirtronPred web server is available at

http://bioinfo.icgeb.res.in/mirtronPred.

Copyright © 2012 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ygeno.2012.04.002

PMID: 22546559 [Indexed for MEDLINE]

1426. Infect Genet Evol. 2012 Jun;12(4):789-97. doi: 10.1016/j.meegid.2012.02.010. Epub

2012 Mar 3.

TB-Lineage: an online tool for classification and analysis of strains of

Mycobacterium tuberculosis complex.

Shabbeer A(1), Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, Bennett

KP.

Author information:

(1)Computer Science Dept., Rensselaer Polytechnic Institute, Troy, NY, USA.

shabba@cs.rpi.edu

This paper formulates a set of rules to classify genotypes of the Mycobacterium

tuberculosis complex (MTBC) into major lineages using spoligotypes and MIRU-VNTR

results. The rules synthesize prior literature that characterizes lineages by

spacer deletions and variations in the number of repeats seen at locus MIRU24

(alias VNTR2687). A tool that efficiently and accurately implements this rule

base is now freely available at http://tbinsight.cs.rpi.edu/run\_tb\_lineage.html.

When MIRU24 data is not available, the system utilizes predictions made by a

Naïve Bayes classifier based on spoligotype data. This website also provides a

tool to generate spoligoforests in order to visualize the genetic diversity and

relatedness of genotypes and their associated lineages. A detailed analysis of

the application of these tools on a dataset collected by the CDC consisting of

3198 distinct spoligotypes and 5430 distinct MIRU-VNTR types from 37,066 clinical

isolates is presented. The tools were also tested on four other independent

datasets. The accuracy of automated classification using both spoligotypes and

MIRU24 is >99%, and using spoligotypes alone is >95%. This online rule-based

classification technique in conjunction with genotype visualization provides a

practical tool that supports surveillance of TB transmission trends and molecular

epidemiological studies.

Copyright © 2012 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.meegid.2012.02.010

PMID: 22406225 [Indexed for MEDLINE]

1427. J Appl Crystallogr. 2012 Jun 1;45(Pt 3):598-602. Epub 2012 May 16.

PROSPERO: online prediction of crystallographic success from experimental results

and sequence.

Zucker FH(1), Kim HY, Merritt EA.

Author information:

(1)Biomolecular Structure Center, Department of Biochemistry, University of

Washington, Seattle, WA 98195-7742, USA.

The growth of diffracting crystals from purified proteins is often a major

bottleneck in determining structures of biological and medical interest. The

PROSPERO web server, http://skuld.bmsc.washington.edu/prospero, is intended both

to provide a means of organizing the potentially large numbers of experimental

characterizations measured from such proteins, and to provide useful guidance for

structural biologists who have succeeded in purifying their target protein but

have reached an impasse in the difficult and poorly understood process of turning

purified protein into well diffracting crystals. These researchers need to decide

which of many possible rescue options are worth pursuing, given finite resources.

This choice is even more crucial when attempting to solve high-priority but

relatively difficult structures of eukaryotic proteins. The site currently uses

the HyGX1 predictor, which was trained and validated on protein samples from

pathogenic protozoa (eukaryotes) using results from six types of experiment.

PROSPERO allows users to store, analyze and display multiple results for each

sample, to group samples into projects, and to share results and predictions with

collaborators.

DOI: 10.1107/S002188981201775X

PMCID: PMC3359727

PMID: 22675232

1428. J Bioinform Comput Biol. 2012 Jun;10(3):1242010. doi: 10.1142/S0219720012420103.

Using rigidity analysis to probe mutation-induced structural changes in proteins.

Jagodzinski F(1), Hardy J, Streinu I.

Author information:

(1)Department of Computer Science, 140 Governors Drive, University of

Massachusetts Amherst, Amherst, MA 01002, USA.

Predicting the effect of a single amino acid substitution on the stability of a

protein structure is a fundamental task in macromolecular modeling. It has

relevance to drug design and understanding of disease-causing protein variants.

We present KINARI-Mutagen, a web server for performing in silico mutation

experiments on protein structures from the Protein Data Bank. Our

rigidity-theoretical approach permits fast evaluation of the effects of mutations

that may not be easy to perform in vitro, because it is not always possible to

express a protein with a specific amino acid substitution. We use KINARI-Mutagen

to identify critical residues, and we show that our predictions correlate with

destabilizing mutations to glycine. In two in-depth case studies we show that the

mutated residues identified by KINARI-Mutagen as critical correlate with

experimental data, and would not have been identified by other methods such as

Solvent Accessible Surface Area measurements or residue ranking by contributions

to stabilizing interactions. We also generate 48 mutants for 14 proteins, and

compare our rigidity-based results against experimental mutation stability data.

KINARI-Mutagen is available at http://kinari.cs.umass.edu.

DOI: 10.1142/S0219720012420103

PMID: 22809386 [Indexed for MEDLINE]

1429. Mol Cell Proteomics. 2012 Jun;11(6):M111.015974. doi: 10.1074/mcp.M111.015974.

Epub 2012 Feb 7.

msCompare: a framework for quantitative analysis of label-free LC-MS data for

comparative candidate biomarker studies.

Hoekman B(1), Breitling R, Suits F, Bischoff R, Horvatovich P.

Author information:

(1)Department of Pharmacy, University of Groningen, 9713 AV Groningen, The

Netherlands.

Data processing forms an integral part of biomarker discovery and contributes

significantly to the ultimate result. To compare and evaluate various publicly

available open source label-free data processing workflows, we developed

msCompare, a modular framework that allows the arbitrary combination of different

feature detection/quantification and alignment/matching algorithms in conjunction

with a novel scoring method to evaluate their overall performance. We used

msCompare to assess the performance of workflows built from modules of publicly

available data processing packages such as SuperHirn, OpenMS, and MZmine and our

in-house developed modules on peptide-spiked urine and trypsin-digested

cerebrospinal fluid (CSF) samples. We found that the quality of results varied

greatly among workflows, and interestingly, heterogeneous combinations of

algorithms often performed better than the homogenous workflows. Our scoring

method showed that the union of feature matrices of different workflows

outperformed the original homogenous workflows in some cases. msCompare is open

source software (https://trac.nbic.nl/mscompare), and we provide a web-based data

processing service for our framework by integration into the Galaxy server of the

Netherlands Bioinformatics Center (http://galaxy.nbic.nl/galaxy) to allow

scientists to determine which combination of modules provides the most accurate

processing for their particular LC-MS data sets.

DOI: 10.1074/mcp.M111.015974

PMCID: PMC3433919

PMID: 22318370 [Indexed for MEDLINE]

1430. Plant J. 2012 Jun;70(5):891-901. doi: 10.1111/j.1365-313X.2012.04922.x. Epub 2012

Mar 8.

SoMART: a web server for plant miRNA, tasiRNA and target gene analysis.

Li F(1), Orban R, Baker B.

Author information:

(1)Department of Plant and Microbial Biology, Plant Gene Expression Center,

University of California, Berkeley, USA. lifeng@berkeley.edu

Plant microRNAs (miRNAs) and trans-acting small interfering RNAs (tasiRNAs) play

important roles in a variety of biological processes. Bioinformatics prediction

and small RNA (sRNA) cloning are the most important approaches for identification

of miRNAs and tasiRNAs and their targets. However, these approaches are not

readily accessible to every researcher. Here we present SoMART, a web server for

miRNA/tasiRNA analysis resources and tools, which is designed for researchers who

are interested in identifying miRNAs or tasiRNAs that potentially regulate genes

of interest. The server includes four sets of tools: 'Slicer detector' for

detecting sRNAs targeting input genes, 'dRNA mapper' for detecting degradome

(d)RNA products derived from input genes, 'PreMIR detector' for identifying miRNA

precursors (MIRs) or tasiRNA precursor (TASs) of input sRNAs, and 'sRNA mapper'

for mapping sRNAs onto input genes. We also developed a dRNA-seq protocol to

achieve longer dRNA reads for better characterization of miRNA precursors by dRNA

mapper. To validate the server function and robustness, we installed sRNA, dRNA

and collected genomic DNA or transcriptome databases from Arabidopsis and

solanaceous plants, and characterized miR172-mediated regulation of the APETALA2

gene in potato (Solanum tuberosum) and demonstrated conservation of

MIR390-triggered TAS3 in tomato (Solanum lycopersicum). More importantly, we

predicted the existence of MIR482-triggered TAS5 in tomato. We further tested and

confirmed the efficiency and accuracy of the server by analyses of 21 validated

miRNA targets and 115 miRNA precursors in Arabidopsis thaliana. SoMART is

available at http://somart.ist.berkeley.edu.

Published 2012. This article is a US Government work and is in the public domain

in the USA.

DOI: 10.1111/j.1365-313X.2012.04922.x

PMID: 22268718 [Indexed for MEDLINE]

1431. Protein Pept Lett. 2012 Jun 1;19(6):644-51.

Using protein-protein interaction network information to predict the subcellular

locations of proteins in budding yeast.

Hu LL(1), Feng KY, Cai YD, Chou KC.

Author information:

(1)Institute of Systems Biology, Shanghai University, Shanghai, China.

The information of protein subcellular localization is vitally important for

in-depth understanding the intricate pathways that regulate biological processes

at the cellular level. With the rapidly increasing number of newly found protein

sequence in the Post-Genomic Age, many automated methods have been developed

attempting to help annotate their subcellular locations in a timely manner.

However, very few of them were developed using the protein-protein interaction

(PPI) network information. In this paper, we have introduced a new concept called

"tethering potential" by which the PPI information can be effectively fused into

the formulation for protein samples. Based on such a network frame, a new

predictor called Yeast-PLoc has been developed for identifying budding yeast

proteins among their 19 subcellular location sites. Meanwhile, a purely

sequence-based approach, called the "hybrid-property" method, is integrated into

Yeast-PLoc as a fall-back to deal with those proteins without sufficient PPI

information. The overall success rate by the jackknife test on the 4,683 yeast

proteins in the training dataset was 70.25%. Furthermore, it was shown that the

success rate by Yeast- PLoc on an independent dataset was remarkably higher than

those by some other existing predictors, indicating that the current approach by

incorporating the PPI information is quite promising. As a user-friendly

web-server, Yeast-PLoc is freely accessible at http://yeastloc.biosino.org/.

PMID: 22519536 [Indexed for MEDLINE]

1432. Toxicol Appl Pharmacol. 2012 Jun 1;261(2):142-53. doi:

10.1016/j.taap.2012.03.018. Epub 2012 Apr 4.

VirtualToxLab - a platform for estimating the toxic potential of drugs, chemicals

and natural products.

Vedani A(1), Dobler M, Smieško M.

Author information:

(1)Biographics Laboratory 3R, Klingelbergstrasse 50, 4056 Basel, Switzerland.

angelo.vedani@unibas.ch

The VirtualToxLab is an in silico technology for estimating the toxic potential

(endocrine and metabolic disruption, some aspects of carcinogenicity and

cardiotoxicity) of drugs, chemicals and natural products. The technology is based

on an automated protocol that simulates and quantifies the binding of small

molecules towards a series of proteins, known or suspected to trigger adverse

effects. The toxic potential, a non-linear function ranging from 0.0 (none) to

1.0 (extreme), is derived from the individual binding affinities of a compound

towards currently 16 target proteins: 10 nuclear receptors (androgen, estrogen α,

estrogen β, glucocorticoid, liver X, mineralocorticoid, peroxisome

proliferator-activated receptor γ, progesterone, thyroid α, and thyroid β), four

members of the cytochrome P450 enzyme family (1A2, 2C9, 2D6, and 3A4), a

cytosolic transcription factor (aryl hydrocarbon receptor) and a potassium ion

channel (hERG). The interface to the technology allows building and uploading

molecular structures, viewing and downloading results and, most importantly,

rationalizing any prediction at the atomic level by interactively analyzing the

binding mode of a compound with its target protein(s) in real-time 3D. The

VirtualToxLab has been used to predict the toxic potential for over 2500

compounds: the results are posted on http://www.virtualtoxlab.org. The free

platform - the OpenVirtualToxLab - is accessible (in client-server mode) over the

Internet. It is free of charge for universities, governmental agencies,

regulatory bodies and non-profit organizations.

Copyright © 2012 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.taap.2012.03.018

PMID: 22521603 [Indexed for MEDLINE]

1433. BMC Bioinformatics. 2012 May 24;13:111. doi: 10.1186/1471-2105-13-111.

MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins.

Kozlowski LP(1), Bujnicki JM.

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology, ul, Trojdena 4, 02-109, Warsaw, Poland.

BACKGROUND: Intrinsically unstructured proteins (IUPs) lack a well-defined

three-dimensional structure. Some of them may assume a locally stable structure

under specific conditions, e.g. upon interaction with another molecule, while

others function in a permanently unstructured state. The discovery of IUPs

challenged the traditional protein structure paradigm, which stated that a

specific well-defined structure defines the function of the protein. As of

December 2011, approximately 60 methods for computational prediction of protein

disorder from sequence have been made publicly available. They are based on

different approaches, such as utilizing evolutionary information, energy

functions, and various statistical and machine learning methods.

RESULTS: Given the diversity of existing intrinsic disorder prediction methods,

we decided to test whether it is possible to combine them into a more accurate

meta-prediction method. We developed a method based on arbitrarily chosen 13

disorder predictors, in which the final consensus was weighted by the accuracy of

the methods. We have also developed a disorder predictor GSmetaDisorder3D that

used no third-party disorder predictors, but alignments to known protein

structures, reported by the protein fold-recognition methods, to infer the

potentially structured and unstructured regions. Following the success of our

disorder predictors in the CASP8 benchmark, we combined them into a meta-meta

predictor called GSmetaDisorderMD, which was the top scoring method in the

subsequent CASP9 benchmark.

CONCLUSIONS: A series of disorder predictors described in this article is

available as a MetaDisorder web server at

http://iimcb.genesilico.pl/metadisorder/. Results are presented both in an easily

interpretable, interactive mode and in a simple text format suitable for machine

processing.

DOI: 10.1186/1471-2105-13-111

PMCID: PMC3465245

PMID: 22624656 [Indexed for MEDLINE]

1434. Database (Oxford). 2012 May 2;2012:bas024. doi: 10.1093/database/bas024. Print

2012.

Directly e-mailing authors of newly published papers encourages community

curation.

Bunt SM(1), Grumbling GB, Field HI, Marygold SJ, Brown NH, Millburn GH; FlyBase

Consortium.

Collaborators: Gelbart W, Brown N, Cripps R, Kaufman T, Matthews K,

Werner-Washburne M, Adryan B, Costa M, Crosby L, Dirkmaat A, dos Santos G, Emmert

D, Falls K, Field H, Goodman J, Gramates LS, Grumbling G, Marygold S, Matthews B,

McQuilton P, Millburn G, Osumi-Sutherland D, Platero H, Ponting L, Russo S,

Schroeder A, Stefancsik R, St Pierre S, Strelets V, Thurmond J, Tweedie S, Wong

JD, Zhou P, Zytkovicz M.

Author information:

(1)FlyBase, Department of Genetics, University of Cambridge, Downing Street,

Cambridge CB2 3EH, UK.

Much of the data within Model Organism Databases (MODs) comes from manual

curation of the primary research literature. Given limited funding and an

increasing density of published material, a significant challenge facing all MODs

is how to efficiently and effectively prioritize the most relevant research

papers for detailed curation. Here, we report recent improvements to the triaging

process used by FlyBase. We describe an automated method to directly e-mail

corresponding authors of new papers, requesting that they list the genes studied

and indicate ('flag') the types of data described in the paper using an online

tool. Based on the author-assigned flags, papers are then prioritized for

detailed curation and channelled to appropriate curator teams for full data

extraction. The overall response rate has been 44% and the flagging of data types

by authors is sufficiently accurate for effective prioritization of papers. In

summary, we have established a sustainable community curation program, with the

result that FlyBase curators now spend less time triaging and can devote more

effort to the specialized task of detailed data extraction. Database URL:

http://flybase.org/

DOI: 10.1093/database/bas024

PMCID: PMC3342516

PMID: 22554788 [Indexed for MEDLINE]

1435. Amino Acids. 2012 May;42(5):1703-13. doi: 10.1007/s00726-011-0872-8. Epub 2011

Mar 13.

Predicting sub-cellular localization of tRNA synthetases from their primary

structures.

Panwar B(1), Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology (CSIR), Sector 39A,

Chandigarh, India.

Since endo-symbiotic events occur, all genes of mitochondrial aminoacyl tRNA

synthetase (AARS) were lost or transferred from ancestral mitochondrial genome

into the nucleus. The canonical pattern is that both cytosolic and mitochondrial

AARSs coexist in the nuclear genome. In the present scenario all mitochondrial

AARSs are nucleus-encoded, synthesized on cytosolic ribosomes and

post-translationally imported from the cytosol into the mitochondria in

eukaryotic cell. The site-based discrimination between similar types of enzymes

is very challenging because they have almost same physico-chemical properties. It

is very important to predict the sub-cellular location of AARSs, to understand

the mitochondrial protein synthesis. We have analyzed and optimized the

distinguishable patterns between cytosolic and mitochondrial AARSs. Firstly,

support vector machines (SVM)-based modules have been developed using amino acid

and dipeptide compositions and achieved Mathews correlation coefficient (MCC) of

0.82 and 0.73, respectively. Secondly, we have developed SVM modules using

position-specific scoring matrix and achieved the maximum MCC of 0.78. Thirdly,

we developed SVM modules using N-terminal, intermediate residues, C-terminal and

split amino acid composition (SAAC) and achieved MCC of 0.82, 0.70, 0.39 and

0.86, respectively. Finally, a SVM module was developed using selected attributes

of split amino acid composition (SA-SAAC) approach and achieved MCC of 0.92 with

an accuracy of 96.00%. All modules were trained and tested on a non-redundant

data set and evaluated using fivefold cross-validation technique. On the

independent data sets, SA-SAAC based prediction model achieved MCC of 0.95 with

an accuracy of 97.77%. The web-server 'MARSpred' based on above study is

available at http://www.imtech.res.in/raghava/marspred/.

DOI: 10.1007/s00726-011-0872-8

PMID: 21400228 [Indexed for MEDLINE]

1436. Anesth Analg. 2012 May;114(5):947-55. doi: 10.1213/ANE.0b013e31824c4def. Epub

2012 Mar 5.

Binding site and affinity prediction of general anesthetics to protein targets

using docking.

Liu R(1), Perez-Aguilar JM, Liang D, Saven JG.

Author information:

(1)Department of Anesthesiology and Critical Care, Hospital of University of

Pennsylvania, Philadelphia, PA 19104, USA. liur@uphs.upenn.edu

BACKGROUND: The protein targets for general anesthetics remain unclear. A tool to

predict anesthetic binding for potential binding targets is needed. In this

study, we explored whether a computational method, AutoDock, could serve as such

a tool.

METHODS: High-resolution crystal data of water-soluble proteins (cytochrome C,

apoferritin, and human serum albumin), and a membrane protein (a pentameric

ligand-gated ion channel from Gloeobacter violaceus [GLIC]) were used. Isothermal

titration calorimetry (ITC) experiments were performed to determine anesthetic

affinity in solution conditions for apoferritin. Docking calculations were

performed using DockingServer with the Lamarckian genetic algorithm and the Solis

and Wets local search method (http://www.dockingserver.com/web). Twenty general

anesthetics were docked into apoferritin. The predicted binding constants were

compared with those obtained from ITC experiments for potential correlations. In

the case of apoferritin, details of the binding site and their interactions were

compared with recent cocrystallization data. Docking calculations for 6 general

anesthetics currently used in clinical settings (isoflurane, sevoflurane,

desflurane, halothane, propofol, and etomidate) with known 50% effective

concentration (EC(50)) values were also performed in all tested proteins. The

binding constants derived from docking experiments were compared with known

EC(50) values and octanol/water partition coefficients for the 6 general

anesthetics.

RESULTS: All 20 general anesthetics docked unambiguously into the anesthetic

binding site identified in the crystal structure of apoferritin. The binding

constants for 20 anesthetics obtained from the docking calculations correlate

significantly with those obtained from ITC experiments (P = 0.04). In the case of

GLIC, the identified anesthetic binding sites in the crystal structure are among

the docking predicted binding sites, but not the top ranked site. Docking

calculations suggest a most probable binding site located in the extracellular

domain of GLIC. The predicted affinities correlated significantly with the known

EC(50) values for the 6 frequently used anesthetics in GLIC for the site

identified in the experimental crystal data (P = 0.006). However, predicted

affinities in apoferritin, human serum albumin, and cytochrome C did not

correlate with these 6 anesthetics' known experimental EC(50) values. A weak

correlation between the predicted affinities and the octanol/water partition

coefficients was observed for the sites in GLIC.

CONCLUSION: We demonstrated that anesthetic binding sites and relative affinities

can be predicted using docking calculations in an automatic docking server

(AutoDock) for both water-soluble and membrane proteins. Correlation of predicted

affinity and EC(50) for 6 frequently used general anesthetics was only observed

in GLIC, a member of a protein family relevant to anesthetic mechanism.

DOI: 10.1213/ANE.0b013e31824c4def

PMCID: PMC3334420

PMID: 22392968 [Indexed for MEDLINE]

1437. Biochemistry (Mosc). 2012 May;77(5):435-43. doi: 10.1134/S0006297912050033.

Classification of rhodopsin structures by modern methods of structural

bioinformatics.

Novikov GV(1), Sivozhelezov VS, Shebanova AS, Shaitan KV.

Author information:

(1)Institute of Cell Biophysics, Russian Academy of Sciences, ul. Institutskaya

3, 142290 Pushchino, Moscow Region, Russia.

We report a classification of the crystallographic structures of bovine and squid

rhodopsins corresponding to different stages of their photocycles. Using the

resource Protein (Structure) Comparison, Knowledge, Similarity, and Information

server (ProCKSI, http://www.procksi.net/), selected spatial structures were

compared on the basis of classification schemes (dendrograms). To compare the

spatial structures of transmembrane proteins, optimal consensus was developed

from methods implemented in ProCKSI. Structures were also clustered using

principal component analysis, resulting in good agreement with the classification

based on the ProCKSI consensus method. Analysis of the results revealed the basic

movements of individual transmembrane domains of these proteins that we were able

to relate to different stages of the photoactivation of rhodopsin. A combination

of methods identified in this study can be used as an up-to-date analytical tool

to study the conformational dynamics of membrane receptors.

DOI: 10.1134/S0006297912050033

PMID: 22813584 [Indexed for MEDLINE]

1438. Bioinformatics. 2012 May 1;28(9):1239-45. doi: 10.1093/bioinformatics/bts119.

Epub 2012 Mar 13.

PINALOG: a novel approach to align protein interaction networks--implications for

complex detection and function prediction.

Phan HT(1), Sternberg MJ.

Author information:

(1)Division of Molecular Biosciences, Faculty of Natural Sciences, Imperial

College, London, UK.

MOTIVATION: Analysis of protein-protein interaction networks (PPINs) at the

system level has become increasingly important in understanding biological

processes. Comparison of the interactomes of different species not only provides

a better understanding of species evolution but also helps with detecting

conserved functional components and in function prediction. Method and

RESULTS: Here we report a PPIN alignment method, called PINALOG, which combines

information from protein sequence, function and network topology. Alignment of

human and yeast PPINs reveals several conserved subnetworks between them that

participate in similar biological processes, notably the proteasome and

transcription related processes. PINALOG has been tested for its power in protein

complex prediction as well as function prediction. Comparison with PSI-BLAST in

predicting protein function in the twilight zone also shows that PINALOG is

valuable in predicting protein function.

AVAILABILITY AND IMPLEMENTATION: The PINALOG web-server is freely available from

http://www.sbg.bio.ic.ac.uk/~pinalog. The PINALOG program and associated data are

available from the Download section of the web-server.

CONTACT: m.sternberg@imperial.ac.uk

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/bts119

PMCID: PMC3338015

PMID: 22419782 [Indexed for MEDLINE]

1439. IEEE Trans Inf Technol Biomed. 2012 May;16(3):356-64. doi:

10.1109/TITB.2011.2176497. Epub 2011 Nov 18.

A RESTful image gateway for multiple medical image repositories.

Valente F(1), Viana-Ferreira C, Costa C, Oliveira JL.

Author information:

(1)IEETA, Universidade de Aveiro, Aveiro, Portugal. fmvalente@ua.pt

Mobile technologies are increasingly important components in telemedicine systems

and are becoming powerful decision support tools. Universal access to data may

already be achieved by resorting to the latest generation of tablet devices and

smartphones. However, the protocols employed for communicating with image

repositories are not suited to exchange data with mobile devices. In this paper,

we present an extensible approach to solving the problem of querying and

delivering data in a format that is suitable for the bandwidth and graphic

capacities of mobile devices. We describe a three-tiered component-based gateway

that acts as an intermediary between medical applications and a number of Picture

Archiving and Communication Systems (PACS). The interface with the gateway is

accomplished using Hypertext Transfer Protocol (HTTP) requests following a

Representational State Transfer (REST) methodology, which relieves developers

from dealing with complex medical imaging protocols and allows the processing of

data on the server side.

DOI: 10.1109/TITB.2011.2176497

PMID: 22113810 [Indexed for MEDLINE]

1440. J Rheumatol. 2012 May;39(5):1088-94. doi: 10.3899/jrheum.111030. Epub 2012 Feb

15.

Assessing the performance of the Birmingham Vasculitis Activity Score at

diagnosis for children with antineutrophil cytoplasmic antibody-associated

vasculitis in A Registry for Childhood Vasculitis (ARChiVe).

Morishita K(1), Li SC, Muscal E, Spalding S, Guzman J, Uribe A, Abramson L,

Baszis K, Benseler S, Bowyer S, Campillo S, Chira P, Hersh AO, Higgins G,

Eberhard A, Ede K, Imundo L, Jung L, Kim S, Kingsbury DJ, Klein-Gitelman M,

Lawson EF, Lovell DJ, Mason T, McCurdy D, Nanda K, Nassi L, O'Neil KM, Rabinovich

E, Ramsey SE, Reiff A, Rosenkranz M, Schikler K, Stevens A, Wahezi D, Cabral DA;

ARChiVe Investigators Network.

Collaborators: Sarmiento A, Espinosa V, Houghton K, Petty R, Tucker L, Turvey S,

Brooks EB, Robinson A, Singer NG, Ilowite NT, Dedeoglu F, Fuhlbrigge R, Hazen M,

Son MB, Sundel R, Brown D, Shaham B, Adams M, Valentini R, Hirsh R, Kietz D,

Rosen P, Torok K, Pachman L, Brunner H, Griffin T, Grom A, Zeft A, Hashkes P,

Eichenfield A, Ardoin S, Schanberg L, Laxer R, Schneider R, Huber AM, Lang BA,

Stringer E, Haines K, Kimura Y, Weiss J, Lee T, Balboni I, Bromberg R, Cidon M,

Frankovich J, Gerstbacher D, Hsu JJ, Park JL, Sandborg C, Song S, Reed A,

Magalnick M, Ramirez A, Shishov M, Ballinger S, Klausmeier T, White A, Emery H,

Hayward K, Ringold S, Shaw E, Turner J, Wallace C, Myones BL, Cartwright V,

Chédeville G, Duffy C, Duffy K, Scuccimarri R, von Scheven E, Jarvis J, Punaro M,

Pascual V, Bonsack J, Prahalad S, Abramson L.

Author information:

(1)British Columbia Children’s Hospital, Vancouver, British Columbia, Canada.

OBJECTIVE: There are no validated tools for measuring disease activity in

pediatric vasculitis. The Birmingham Vasculitis Activity Score (BVAS) is a valid

disease activity tool in adult vasculitis. Version 3 (BVAS v.3) correlates well

with physician's global assessment (PGA), treatment decision, and C-reactive

protein in adults. The utility of BVAS v.3 in pediatric vasculitis is not known.

We assessed the association of BVAS v.3 scores with PGA, treatment decision, and

erythrocyte sedimentation rate (ESR) at diagnosis in pediatric antineutrophil

cytoplasmic antibody-associated vasculitis (AAV).

METHODS: Children with AAV diagnosed between 2004 and 2010 at all ARChiVe centers

were eligible. BVAS v.3 scores were calculated with a standardized online tool

(www.vasculitis.org). Spearman's rank correlation coefficient (r(s)) was used to

test the strength of association between BVAS v.3 and PGA, treatment decision,

and ESR.

RESULTS: A total of 152 patients were included. The physician diagnosis of these

patients was predominantly granulomatosis with polyangiitis (n = 99). The median

BVAS v.3 score was 18.0 (range 0-40). The BVAS v.3 correlations were r(s) = 0.379

(95% CI 0.233 to 0.509) with PGA, r(s) = 0.521 (95% CI 0.393 to 0.629) with

treatment decision, and r(s) = 0.403 (95% CI 0.253 to 0.533) with ESR.

CONCLUSION: Applied to children with AAV, BVAS v.3 had a weak correlation with

PGA and moderate correlation with both ESR and treatment decision. Prospective

evaluation of BVAS v.3 and/or pediatric-specific modifications to BVAS v.3 may be

required before it can be formalized as a disease activity assessment tool in

pediatric AAV.

DOI: 10.3899/jrheum.111030

PMID: 22337238 [Indexed for MEDLINE]

1441. RNA. 2012 May;18(5):900-14. doi: 10.1261/rna.029041.111. Epub 2012 Mar 26.

LocARNA-P: accurate boundary prediction and improved detection of structural

RNAs.

Will S(1), Joshi T, Hofacker IL, Stadler PF, Backofen R.

Author information:

(1)Chair for Bioinformatics, Institute of Computer Science,

Albert-Ludwigs-Universität, D-79110 Freiburg, Germany.

Current genomic screens for noncoding RNAs (ncRNAs) predict a large number of

genomic regions containing potential structural ncRNAs. The analysis of these

data requires highly accurate prediction of ncRNA boundaries and discrimination

of promising candidate ncRNAs from weak predictions. Existing methods struggle

with these goals because they rely on sequence-based multiple sequence

alignments, which regularly misalign RNA structure and therefore do not support

identification of structural similarities. To overcome this limitation, we

compute columnwise and global reliabilities of alignments based on sequence and

structure similarity; we refer to these structure-based alignment reliabilities

as STARs. The columnwise STARs of alignments, or STAR profiles, provide a

versatile tool for the manual and automatic analysis of ncRNAs. In particular, we

improve the boundary prediction of the widely used ncRNA gene finder RNAz by a

factor of 3 from a median deviation of 47 to 13 nt. Post-processing RNAz

predictions, LocARNA-P's STAR score allows much stronger discrimination between

true- and false-positive predictions than RNAz's own evaluation. The improved

accuracy, in this scenario increased from AUC 0.71 to AUC 0.87, significantly

reduces the cost of successive analysis steps. The ready-to-use software tool

LocARNA-P produces structure-based multiple RNA alignments with associated

columnwise STARs and predicts ncRNA boundaries. We provide additional results, a

web server for LocARNA/LocARNA-P, and the software package, including

documentation and a pipeline for refining screens for structural ncRNA, at

http://www.bioinf.uni-freiburg.de/Supplements/LocARNA-P/.

DOI: 10.1261/rna.029041.111

PMCID: PMC3334699

PMID: 22450757 [Indexed for MEDLINE]

1442. ACS Synth Biol. 2012 Apr 20;1(4):139-50. doi: 10.1021/sb200019x. Epub 2012 Feb

22.

MAP(2.0)3D: a sequence/structure based server for protein engineering.

Verma R(1), Schwaneberg U, Roccatano D.

Author information:

(1)School of Engineering and Science, Jacobs University Bremen, Campus Ring 1,

28759 Bremen, Germany.

The Mutagenesis Assistant Program (MAP) is a web-based tool to provide

statistical analyses of the mutational biases of directed evolution experiments

on amino acid substitution patterns. MAP analysis assists protein engineers in

the benchmarking of random mutagenesis methods that generate single nucleotide

mutations in a codon. Herein, we describe a completely renewed and improved

version of the MAP server, the MAP(2.0)3D server, which correlates the generated

amino acid substitution patterns to the structural information of the target

protein. This correlation aids in the selection of a more suitable random

mutagenesis method with specific biases on amino acid substitution patterns. In

particular, the new server represents MAP indicators on secondary and tertiary

structure and correlates them to specific structural components such as hydrogen

bonds, hydrophobic contacts, salt bridges, solvent accessibility, and

crystallographic B-factors. Three model proteins (D-amino oxidase, phytase, and

N-acetylneuraminic acid aldolase) are used to illustrate the novel capability of

the server. MAP(2.0)3D server is available publicly at

http://map.jacobs-university.de/map3d.html.

DOI: 10.1021/sb200019x

PMID: 23651115 [Indexed for MEDLINE]

1443. Front Genet. 2012 Apr 18;3:59. doi: 10.3389/fgene.2012.00059. eCollection 2012.

HD-RNAS: An Automated Hierarchical Database of RNA Structures.

Ray SS(1), Halder S, Kaypee S, Bhattacharyya D.

Author information:

(1)Machine Intelligence Unit, Indian Statistical Institute Kolkata, India.

One of the important goals of most biological investigations is to classify and

organize the experimental findings so that they are readily useful for deriving

generalized rules. Although there is a huge amount of information on RNA

structures in PDB, there are redundant files, ambiguous synthetic sequences etc.

Moreover, a systematic hierarchical organization, reflecting RNA classification,

is missing in PDB. In this investigation, we have classified all the available

RNA structures from PDB through a programmatic approach. Hence, it would be now a

simple assignment to regularly update the classification as and when new

structures are released. The classification can further determine (i) a

non-redundant set of RNA structures and (ii) if available, a set of structures of

identical sequence and function, which can highlight structural polymorphism,

ligand-induced conformational alterations etc. Presently, we have classified the

available structures (2095 PDB entries having RNA chain longer than nine

nucleotides solved by X-ray crystallography or NMR spectroscopy) into nine

functional classes. The structures of same function and same source are mostly

seen to be similar with subtle differences depending on their functional

complexation. The web-server is available online at

http://www.saha.ac.in/biop/www/HD-RNAS.html and is updated regularly.

DOI: 10.3389/fgene.2012.00059

PMCID: PMC3329738

PMID: 22529851

1444. Bioinformatics. 2012 Apr 15;28(8):1172-3. doi: 10.1093/bioinformatics/bts095.

Epub 2012 Feb 24.

SiteComp: a server for ligand binding site analysis in protein structures.

Lin Y(1), Yoo S, Sanchez R.

Author information:

(1)Department of Structural and Chemical Biology, Mount Sinai School of Medicine,

New York, NY 10029, USA.

MOTIVATION: Computational characterization of ligand-binding sites in proteins

provides preliminary information for functional annotation, protein design and

ligand optimization. SiteComp implements binding site analysis for comparison of

binding sites, evaluation of residue contribution to binding sites and

identification of sub-sites with distinct molecular interaction properties.

AVAILABILITY AND IMPLEMENTATION: The SiteComp server and tutorials are freely

available at http://sitecomp.sanchezlab.org.

DOI: 10.1093/bioinformatics/bts095

PMCID: PMC3324516

PMID: 22368247 [Indexed for MEDLINE]

1445. Endocr Relat Cancer. 2012 Apr 10;19(2):197-208. doi: 10.1530/ERC-11-0329. Print

2012 Apr.

Implementing an online tool for genome-wide validation of survival-associated

biomarkers in ovarian-cancer using microarray data from 1287 patients.

Gyorffy B(1), Lánczky A, Szállási Z.

Author information:

(1)Research Laboratory of Pediatrics and Nephrology, Hungarian Academy of

Sciences, Budapest, Hungary. gyorffy@kmplot.com

The validation of prognostic biomarkers in large independent patient cohorts is a

major bottleneck in ovarian cancer research. We implemented an online tool to

assess the prognostic value of the expression levels of all microarray-quantified

genes in ovarian cancer patients. First, a database was set up using gene

expression data and survival information of 1287 ovarian cancer patients

downloaded from Gene Expression Omnibus and The Cancer Genome Atlas (Affymetrix

HG-U133A, HG-U133A 2.0, and HG-U133 Plus 2.0 microarrays). After quality control

and normalization, only probes present on all three Affymetrix platforms were

retained (n=22,277). To analyze the prognostic value of the selected gene, we

divided the patients into two groups according to various quantile expressions of

the gene. These groups were then compared using progression-free survival

(n=1090) or overall survival (n=1287). A Kaplan-Meier survival plot was generated

and significance was computed. The tool can be accessed online at

www.kmplot.com/ovar. We used this integrative data analysis tool to validate the

prognostic power of 37 biomarkers identified in the literature. Of these, CA125

(MUC16; P=3.7×10(-5), hazard ratio (HR)=1.4), CDKN1B (P=5.4×10(-5), HR=1.4), KLK6

(P=0.002, HR=0.79), IFNG (P=0.004, HR=0.81), P16 (P=0.02, HR=0.66), and BIRC5

(P=0.00017, HR=0.75) were associated with survival. The combination of several

probe sets can further increase prediction efficiency. In summary, we developed a

global online biomarker validation platform that mines all available microarray

data to assess the prognostic power of 22,277 genes in 1287 ovarian cancer

patients. We specifically used this tool to evaluate the effect of 37 previously

published biomarkers on ovarian cancer prognosis.

DOI: 10.1530/ERC-11-0329

PMID: 22277193 [Indexed for MEDLINE]

1446. Anim Sci J. 2012 Apr;83(4):279-83. doi: 10.1111/j.1740-0929.2011.00952.x. Epub

2011 Sep 12.

SNPpath: characterizing cattle SNPs by enriched pathway terms.

Wang Q(1), Wang M, Yang Y, Pan Y.

Author information:

(1)School of Agriculture and Biology, Shanghai Jiao Tong University, China.

High-density single nucleotide polymorphism (SNP) microarrays have made

large-scale genome-wide association studies (GWAS) and genomic selection (GS)

feasible. Valuable insight into the genetic basis underlying complex polygenic

traits will likely be gained by considering functionally related sets of genes

simultaneously. SNPpath, a suite of computer-generated imagery-based web servers

has been developed to automatically annotate and characterize cattle SNPs by

enriched KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway terms. The

SNPpath allows users to navigate and analysis large SNP sets and is the only web

server currently providing pathway annotations of cattle SNPs in National Center

for Biotechnology Information's dbSNP database and three commercial platforms.

Hence, we describe SNPpath and provide details of the query options, as well as

biological examples of use. The SNPpath may be favorable for the analysis of

combining SNP association analysis with pathway-driven gene set enrichment

analysis and is freely available at http://klab.sjtu.edu.cn/SNPpath.

© 2011 The Authors. Animal Science Journal © 2011 Japanese Society of Animal

Science.

DOI: 10.1111/j.1740-0929.2011.00952.x

PMID: 22515686 [Indexed for MEDLINE]

1447. Bioinformatics. 2012 Apr 1;28(7):1040-1. doi: 10.1093/bioinformatics/bts076. Epub

2012 Feb 15.

DOMIRE: a web server for identifying structural domains and their neighbors in

proteins.

Samson F(1), Shrager R, Tai CH, Sam V, Lee B, Munson PJ, Gibrat JF, Garnier J.

Author information:

(1)Institut National de la Recherche Agronomique, UR1077, Unité Mathématique,

Informatique et Génome, 78350 Jouy-en-Josas, France.

SUMMARY: The DOMIRE web server implements a novel, automatic, protein structural

domain assignment procedure based on 3D substructures of the query protein which

are also found within structures of a non-redundant protein database. These

common 3D substructures are transformed into a co-occurrence matrix that offers a

global view of the protein domain organization. Three different algorithms are

employed to define structural domain boundaries from this co-occurrence matrix.

For each query, a list of structural neighbors and their alignments are provided.

DOMIRE, by displaying the protein structural domain organization, can be a useful

tool for defining protein common cores and for unravelling the evolutionary

relationship between different proteins.

AVAILABILITY: http://genome.jouy.inra.fr/domire

CONTACT: jean.garnier@jouy.inra.fr.

DOI: 10.1093/bioinformatics/bts076

PMCID: PMC3315711

PMID: 22345617 [Indexed for MEDLINE]

1448. Bioinformatics. 2012 Apr 1;28(7):914-20. doi: 10.1093/bioinformatics/bts078. Epub

2012 Feb 10.

WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based

filtering.

Liu Z(1), Abbas A, Jing BY, Gao X.

Author information:

(1)The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen

361000, China.

MOTIVATION: Nuclear magnetic resonance (NMR) has been widely used as a powerful

tool to determine the 3D structures of proteins in vivo. However, the

post-spectra processing stage of NMR structure determination usually involves a

tremendous amount of time and expert knowledge, which includes peak picking,

chemical shift assignment and structure calculation steps. Detecting accurate

peaks from the NMR spectra is a prerequisite for all following steps, and thus

remains a key problem in automatic NMR structure determination.

RESULTS: We introduce WaVPeak, a fully automatic peak detection method. WaVPeak

first smoothes the given NMR spectrum by wavelets. The peaks are then identified

as the local maxima. The false positive peaks are filtered out efficiently by

considering the volume of the peaks. WaVPeak has two major advantages over the

state-of-the-art peak-picking methods. First, through wavelet-based smoothing,

WaVPeak does not eliminate any data point in the spectra. Therefore, WaVPeak is

able to detect weak peaks that are embedded in the noise level. NMR

spectroscopists need the most help isolating these weak peaks. Second, WaVPeak

estimates the volume of the peaks to filter the false positives. This is more

reliable than intensity-based filters that are widely used in existing methods.

We evaluate the performance of WaVPeak on the benchmark set proposed by PICKY

(Alipanahi et al., 2009), one of the most accurate methods in the literature. The

dataset comprises 32 2D and 3D spectra from eight different proteins.

Experimental results demonstrate that WaVPeak achieves an average of 96%, 91%,

88%, 76% and 85% recall on (15)N-HSQC, HNCO, HNCA, HNCACB and CBCA(CO)NH,

respectively. When the same number of peaks are considered, WaVPeak significantly

outperforms PICKY.

AVAILABILITY: WaVPeak is an open source program. The source code and two test

spectra of WaVPeak are available at

http://faculty.kaust.edu.sa/sites/xingao/Pages/Publications.aspx. The online

server is under construction.

CONTACT: statliuzhi@xmu.edu.cn; ahmed.abbas@kaust.edu.sa; majing@ust.hk;

xin.gao@kaust.edu.sa.

DOI: 10.1093/bioinformatics/bts078

PMCID: PMC3315717

PMID: 22328784 [Indexed for MEDLINE]

1449. Bioinformatics. 2012 Apr 1;28(7):1031-2. doi: 10.1093/bioinformatics/bts074. Epub

2012 Feb 10.

Spliceman--a computational web server that predicts sequence variations in

pre-mRNA splicing.

Lim KH(1), Fairbrother WG.

Author information:

(1)Department of Molecular Biology, Cellular Biology and Biochemistry, Brown

University, Providence, RI 02903, USA.

SUMMARY: It was previously demonstrated that splicing elements are positional

dependent. We exploited this relationship between location and function by

comparing positional distributions between all possible 4096 hexamers around a

database of human splice sites. The distance measure used in this study found

point mutations that produced higher distances disrupted splicing, whereas point

mutations with smaller distances generally had no effect on splicing. Reasoning

the idea that functional splicing elements have signature positional

distributions around constitutively spliced exons, we introduce Spliceman-an

online tool that predicts how likely distant mutations around annotated splice

sites were to disrupt splicing. Spliceman takes a set of DNA sequences with point

mutations and returns a ranked list to predict the effects of point mutations on

pre-mRNA splicing. The current implementation included the analyses of 11

genomes: human, chimp, rhesus, mouse, rat, dog, cat, chicken, guinea pig, frog

and zebrafish.

AVAILABILITY: Freely available on the web at

http://fairbrother.biomed.brown.edu/spliceman/

CONTACT: fairbrother@brown.edu.

DOI: 10.1093/bioinformatics/bts074

PMCID: PMC3315715

PMID: 22328782 [Indexed for MEDLINE]

1450. Bioinformatics. 2012 Apr 1;28(7):1028-30. doi: 10.1093/bioinformatics/bts062.

Epub 2012 Jan 31.

WegoLoc: accurate prediction of protein subcellular localization using weighted

Gene Ontology terms.

Chi SM(1), Nam D.

Author information:

(1)School of Computer Science and Engineering, Kyungsung University, Nam-gu,

Suyoung-ro 309, Pusan, South Korea. smchiks@ks.ac.kr

SUMMARY: We present an accurate and fast web server, WegoLoc for predicting

subcellular localization of proteins based on sequence similarity and weighted

Gene Ontology (GO) information. A term weighting method in the text

categorization process is applied to GO terms for a support vector machine

classifier. As a result, WegoLoc surpasses the state-of-the-art methods for

previously used test datasets. WegoLoc supports three eukaryotic kingdoms

(animals, fungi and plants) and provides human-specific analysis, and covers

several sets of cellular locations. In addition, WegoLoc provides (i) multiple

possible localizations of input protein(s) as well as their corresponding

probability scores, (ii) weights of GO terms representing the contribution of

each GO term in the prediction, and (iii) a BLAST E-value for the best hit with

GO terms. If the similarity score does not meet a given threshold, an amino acid

composition-based prediction is applied as a backup method.

AVAILABILITY: WegoLoc and User's guide are freely available at the website

http://www.btool.org/WegoLoc

CONTACT: smchiks@ks.ac.kr; dougnam@unist.ac.kr

SUPPLEMENTARY INFORMATION: Supplementary data is available at

http://www.btool.org/WegoLoc.

DOI: 10.1093/bioinformatics/bts062

PMID: 22296788 [Indexed for MEDLINE]

1451. Environ Manage. 2012 Apr;49(4):816-32. doi: 10.1007/s00267-012-9818-5. Epub 2012

Feb 28.

An integrated WebGIS framework for volunteered geographic information and social

media in soil and water conservation.

Werts JD(1), Mikhailova EA, Post CJ, Sharp JL.

Author information:

(1)Department of Forestry and Natural Resources, Clemson University, 261 Lehotsky

Hall, Clemson, SC 29634, USA.

Volunteered geographic information and social networking in a WebGIS has the

potential to increase public participation in soil and water conservation,

promote environmental awareness and change, and provide timely data that may be

otherwise unavailable to policymakers in soil and water conservation management.

The objectives of this study were: (1) to develop a framework for combining

current technologies, computing advances, data sources, and social media; and (2)

develop and test an online web mapping interface. The mapping interface

integrates Microsoft Silverlight, Bing Maps, ArcGIS Server, Google Picasa Web

Albums Data API, RSS, Google Analytics, and Facebook to create a rich user

experience. The website allows the public to upload photos and attributes of

their own subdivisions or sites they have identified and explore other

submissions. The website was made available to the public in early February 2011

at http://www.AbandonedDevelopments.com and evaluated for its potential long-term

success in a pilot study.

DOI: 10.1007/s00267-012-9818-5

PMID: 22371128 [Indexed for MEDLINE]

1452. Genomics. 2012 Apr;99(4):195-201. doi: 10.1016/j.ygeno.2012.01.008. Epub 2012 Feb

3.

C16S - a Hidden Markov Model based algorithm for taxonomic classification of 16S

rRNA gene sequences.

Ghosh TS(1), Gajjalla P, Mohammed MH, Mande SS.

Author information:

(1)Bio-sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services

Limited, 1 Software Units Layout, Madhapur, Hyderabad, 500081, Andhra Pradesh,

India. tarini@atc.tcs.com

Recent advances in high throughput sequencing technologies and concurrent

refinements in 16S rDNA isolation techniques have facilitated the rapid

extraction and sequencing of 16S rDNA content of microbial communities. The

taxonomic affiliation of these 16S rDNA fragments is subsequently obtained using

either BLAST-based or word frequency based approaches. However, the

classification accuracy of such methods is observed to be limited in typical

metagenomic scenarios, wherein a majority of organisms are hitherto unknown. In

this study, we present a 16S rDNA classification algorithm, called C16S, that

uses genus-specific Hidden Markov Models for taxonomic classification of 16S rDNA

sequences. Results obtained using C16S have been compared with the widely used

RDP classifier. The performance of C16S algorithm was observed to be consistently

higher than the RDP classifier. In some scenarios, this increase in accuracy is

as high as 34%. A web-server for the C16S algorithm is available at

http://metagenomics.atc.tcs.com/C16S/.

Copyright © 2012 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.ygeno.2012.01.008

PMID: 22326741 [Indexed for MEDLINE]

1453. J Am Coll Surg. 2012 Apr;214(4):608-17; discussion 617-9. doi:

10.1016/j.jamcollsurg.2011.12.027. Epub 2012 Feb 17.

A novel and accurate computer model of melanoma prognosis for patients staged by

sentinel lymph node biopsy: comparison with the American Joint Committee on

Cancer model.

Callender GG(1), Gershenwald JE, Egger ME, Scoggins CR, Martin RC 2nd, Schacherer

CW, Edwards MJ, Urist MM, Ross MI, Stromberg AJ, McMasters KM.

Author information:

(1)Department of Surgery, University of Louisville, Louisville, KY 40202, USA.

BACKGROUND: We found that a computer model developed by the American Joint

Committee on Cancer (AJCC) melanoma staging committee had limitations for

predicting prognosis of patients staged by sentinel lymph node (SLN) biopsy. We

sought to develop a model that more accurately predicts prognosis in this

population.

STUDY DESIGN: Using a data set obtained from a prospective multi-institutional

study of 2,507 patients with clinically node-negative melanomas ≥1.0 mm Breslow

thickness, we developed a prognostic model using a Cox regression formula

incorporating a number of significant clinicopathologic factors. The AJCC model

and our model were used to predict 5-year survival from this test data set. The

concordance correlation coefficient (CCC) was determined and chi-square tests

were performed. Our new prognostic model was validated using an independent data

set of 1,001 patients.

RESULTS: Using the test data set, the CCC for the AJCC model was 0.875;

chi-square tests demonstrated statistically significant differences between

observed and predicted survivals for numerous clinicopathologic factors. The CCC

for our model was 0.976 and none of the chi-square tests was statistically

significant. Our model performed similarly well in SLN-negative patients (CCC

0.929) and SLN-positive patients (CCC 0.889). The AJCC model performed well in

SLN-negative patients (CCC 0.854), but not in SLN-positive patients (CCC 0.626).

Using the validation data set, similar findings were obtained.

CONCLUSIONS: Our prognostic model provides superior survival estimates compared

with the AJCC model for patients undergoing SLN biopsy. This online tool is

available at www.melanomacalculator.com, and will provide important information

that can be used to guide adjuvant therapy decisions and stratification in

clinical trials.

Copyright © 2012 American College of Surgeons. Published by Elsevier Inc. All

rights reserved.

DOI: 10.1016/j.jamcollsurg.2011.12.027

PMID: 22342785 [Indexed for MEDLINE]

1454. J Biomol NMR. 2012 Apr;52(4):289-302. doi: 10.1007/s10858-012-9603-z. Epub 2012

Feb 23.

RNA-PAIRS: RNA probabilistic assignment of imino resonance shifts.

Bahrami A(1), Clos LJ 2nd, Markley JL, Butcher SE, Eghbalnia HR.

Author information:

(1)National Magnetic Resonance Facility at Madison, Madison, WI, USA.

The significant biological role of RNA has further highlighted the need for

improving the accuracy, efficiency and the reach of methods for investigating RNA

structure and function. Nuclear magnetic resonance (NMR) spectroscopy is vital to

furthering the goals of RNA structural biology because of its distinctive

capabilities. However, the dispersion pattern in the NMR spectra of RNA makes

automated resonance assignment, a key step in NMR investigation of biomolecules,

remarkably challenging. Herein we present RNA Probabilistic Assignment of Imino

Resonance Shifts (RNA-PAIRS), a method for the automated assignment of RNA imino

resonances with synchronized verification and correction of predicted secondary

structure. RNA-PAIRS represents an advance in modeling the assignment paradigm

because it seeds the probabilistic network for assignment with experimental NMR

data, and predicted RNA secondary structure, simultaneously and from the start.

Subsequently, RNA-PAIRS sets in motion a dynamic network that reverberates

between predictions and experimental evidence in order to reconcile and rectify

resonance assignments and secondary structure information. The procedure is

halted when assignments and base-parings are deemed to be most consistent with

observed crosspeaks. The current implementation of RNA-PAIRS uses an initial peak

list derived from proton-nitrogen heteronuclear multiple quantum correlation

((1)H-(15)N 2D HMQC) and proton-proton nuclear Overhauser enhancement

spectroscopy ((1)H-(1)H 2D NOESY) experiments. We have evaluated the performance

of RNA-PAIRS by using it to analyze NMR datasets from 26 previously studied RNAs,

including a 111-nucleotide complex. For moderately sized RNA molecules, and over

a range of comparatively complex structural motifs, the average assignment

accuracy exceeds 90%, while the average base pair prediction accuracy exceeded

93%. RNA-PAIRS yielded accurate assignments and base pairings consistent with

imino resonances for a majority of the NMR resonances, even when the initial

predictions are only modestly accurate. RNA-PAIRS is available as a public

web-server at http://pine.nmrfam.wisc.edu/RNA/.

DOI: 10.1007/s10858-012-9603-z

PMCID: PMC3480180

PMID: 22359049 [Indexed for MEDLINE]

1455. OMICS. 2012 Apr;16(4):168-77. doi: 10.1089/omi.2011.0115. Epub 2012 Mar 20.

Semirna: searching for plant miRNAs using target sequences.

Muñoz-Mérida A(1), Perkins JR, Viguera E, Thode G, Bejarano ER, Pérez-Pulido AJ.

Author information:

(1)Departamento de Arquitectura de Computadores, Universidad de Málaga, Spain.

Many plant genomes are already known, and new ones are being sequenced every

year. The next step for researchers is to identify all of the functional elements

in these genomes, including the important class of functional elements known as

microRNAs (miRNAs), which are involved in posttranscriptional regulatory

pathways. However, computational tools for predicting new plant miRNAs are

limited, and there is a particular need for tools that can be used easily by

laboratory researchers. We present semirna, a new tool for predicting miRNAs in

plant genomes, available as a Web server. This tool takes a putative target

sequence such as a messenger RNA (mRNA) as input, and allows users to search for

miRNAs that target this sequence. It can also be used to determine whether small

RNA sequences from massive sequencing analysis represent true miRNAs and to

search for miRNAs in new genomes using homology. Semirna has shown a high level

of accuracy using various test sets, and gives users the ability to search for

miRNAs with several different adjustable parameters. Semirna, a user-friendly and

intuitive Web server for predicting miRNA sequences, can be reached at

http://www.bioinfocabd.upo.es/semirna/ . It is useful for researchers searching

for miRNAs involved in particular pathways, as well as those searching for miRNAs

in newly sequenced genomes.

DOI: 10.1089/omi.2011.0115

PMID: 22433074 [Indexed for MEDLINE]

1456. BMC Bioinformatics. 2012 Mar 28;13 Suppl 4:S14. doi: 10.1186/1471-2105-13-S4-S14.

Argot2: a large scale function prediction tool relying on semantic similarity of

weighted Gene Ontology terms.

Falda M(1), Toppo S, Pescarolo A, Lavezzo E, Di Camillo B, Facchinetti A, Cilia

E, Velasco R, Fontana P.

Author information:

(1)Department of Molecular Medicine, University of Padova, via U. Bassi 58/B,

35121, Padova, Italy. marco.falda@unipd.it

BACKGROUND: Predicting protein function has become increasingly demanding in the

era of next generation sequencing technology. The task to assign a

curator-reviewed function to every single sequence is impracticable.

Bioinformatics tools, easy to use and able to provide automatic and reliable

annotations at a genomic scale, are necessary and urgent. In this scenario, the

Gene Ontology has provided the means to standardize the annotation classification

with a structured vocabulary which can be easily exploited by computational

methods.

RESULTS: Argot2 is a web-based function prediction tool able to annotate nucleic

or protein sequences from small datasets up to entire genomes. It accepts as

input a list of sequences in FASTA format, which are processed using BLAST and

HMMER searches vs UniProKB and Pfam databases respectively; these sequences are

then annotated with GO terms retrieved from the UniProtKB-GOA database and the

terms are weighted using the e-values from BLAST and HMMER. The weighted GO terms

are processed according to both their semantic similarity relations described by

the Gene Ontology and their associated score. The algorithm is based on the

original idea developed in a previous tool called Argot. The entire engine has

been completely rewritten to improve both accuracy and computational efficiency,

thus allowing for the annotation of complete genomes.

CONCLUSIONS: The revised algorithm has been already employed and successfully

tested during in-house genome projects of grape and apple, and has proven to have

a high precision and recall in all our benchmark conditions. It has also been

successfully compared with Blast2GO, one of the methods most commonly employed

for sequence annotation. The server is freely accessible at

http://www.medcomp.medicina.unipd.it/Argot2.

DOI: 10.1186/1471-2105-13-S4-S14

PMCID: PMC3314586

PMID: 22536960 [Indexed for MEDLINE]

1457. BMC Bioinformatics. 2012 Mar 28;13 Suppl 4:S1. doi: 10.1186/1471-2105-13-S4-S1.

Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee.

Chang JM(1), Di Tommaso P, Taly JF, Notredame C.

Author information:

(1)Bioinformatics and Genomics program, Centre for Genomic Regulation (CRG) and

UPF, Barcelona 08003, Spain.

BACKGROUND: Transmembrane proteins (TMPs) constitute about 20~30% of all protein

coding genes. The relative lack of experimental structure has so far made it hard

to develop specific alignment methods and the current state of the art (PRALINE™)

only manages to recapitulate 50% of the positions in the reference alignments

available from the BAliBASE2-ref7.

METHODS: We show how homology extension can be adapted and combined with a

consistency based approach in order to significantly improve the multiple

sequence alignment of alpha-helical TMPs. TM-Coffee is a special mode of

PSI-Coffee able to efficiently align TMPs, while using a reduced reference

database for homology extension.

RESULTS: Our benchmarking on BAliBASE2-ref7 alpha-helical TMPs shows a

significant improvement over the most accurate methods such as MSAProbs, Kalign,

PROMALS, MAFFT, ProbCons and PRALINE™. We also estimated the influence of the

database used for homology extension and show that highly non-redundant UniRef

databases can be used to obtain similar results at a significantly reduced

computational cost over full protein databases. TM-Coffee is part of the T-Coffee

package, a web server is also available from http://tcoffee.crg.cat/tmcoffee and

a freeware open source code can be downloaded from

http://www.tcoffee.org/Packages/Stable/Latest.

DOI: 10.1186/1471-2105-13-S4-S1

PMCID: PMC3303701

PMID: 22536955 [Indexed for MEDLINE]

1458. BMC Bioinformatics. 2012 Mar 18;13:41. doi: 10.1186/1471-2105-13-41.

Predicting protein-protein interface residues using local surface structural

similarity.

Jordan RA(1), El-Manzalawy Y, Dobbs D, Honavar V.

Author information:

(1)Department of Computer Science, Iowa State University, Ames, IA 50011, USA.

rjordan@iastate.edu

BACKGROUND: Identification of the residues in protein-protein interaction sites

has a significant impact in problems such as drug discovery. Motivated by the

observation that the set of interface residues of a protein tend to be conserved

even among remote structural homologs, we introduce PrISE, a family of local

structural similarity-based computational methods for predicting protein-protein

interface residues.

RESULTS: We present a novel representation of the surface residues of a protein

in the form of structural elements. Each structural element consists of a central

residue and its surface neighbors. The PrISE family of interface prediction

methods uses a representation of structural elements that captures the atomic

composition and accessible surface area of the residues that make up each

structural element. Each of the members of the PrISE methods identifies for each

structural element in the query protein, a collection of similar structural

elements in its repository of structural elements and weights them according to

their similarity with the structural element of the query protein. PrISEL relies

on the similarity between structural elements (i.e. local structural similarity).

PrISEG relies on the similarity between protein surfaces (i.e. general structural

similarity). PrISEC, combines local structural similarity and general structural

similarity to predict interface residues. These predictors label the central

residue of a structural element in a query protein as an interface residue if a

weighted majority of the structural elements that are similar to it are interface

residues, and as a non-interface residue otherwise. The results of our

experiments using three representative benchmark datasets show that the PrISEC

outperforms PrISEL and PrISEG; and that PrISEC is highly competitive with

state-of-the-art structure-based methods for predicting protein-protein interface

residues. Our comparison of PrISEC with PredUs, a recently developed method for

predicting interface residues of a query protein based on the known interface

residues of its (global) structural homologs, shows that performance superior or

comparable to that of PredUs can be obtained using only local surface structural

similarity. PrISEC is available as a Web server at http://prise.cs.iastate.edu/

CONCLUSIONS: Local surface structural similarity based methods offer a simple,

efficient, and effective approach to predict protein-protein interface residues.

DOI: 10.1186/1471-2105-13-41

PMCID: PMC3386866

PMID: 22424103 [Indexed for MEDLINE]

1459. Bioinformatics. 2012 Mar 15;28(6):893-4. doi: 10.1093/bioinformatics/bts041. Epub

2012 Feb 15.

CytoSaddleSum: a functional enrichment analysis plugin for Cytoscape based on

sum-of-weights scores.

Stojmirovic A(1), Bliskovsky A, Yu YK.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, Bethesda, MD 20894, USA.

CytoSaddleSum provides Cytoscape users with access to the functionality of

SaddleSum, a functional enrichment tool based on sum-of-weight scores. It

operates by querying SaddleSum locally (using the standalone version) or remotely

(through an HTTP request to a web server). The functional enrichment results are

shown as a term relationship network, where nodes represent terms and edges show

term relationships. Furthermore, query results are written as Cytoscape

attributes allowing easy saving, retrieval and integration into network-based

data analysis workflows.

DOI: 10.1093/bioinformatics/bts041

PMCID: PMC3307116

PMID: 22345616 [Indexed for MEDLINE]

1460. BMC Res Notes. 2012 Mar 8;5:130. doi: 10.1186/1756-0500-5-130.

SpiroESTdb: a transcriptome database and online tool for sparganum expressed

sequences tags.

Kim DW(1), Kim DW, Yoo WG, Nam SH, Lee MR, Yang HW, Park J, Lee K, Lee S, Cho SH,

Lee WJ, Park HS, Ju JW.

Author information:

(1)Division of Malaria and Parasitic Diseases, Korea National Institute of

Health, Osong 363-951, Republic of Korea.

BACKGROUND: Sparganum (plerocercoid of Spirometra erinacei) is a parasite that

possesses the remarkable ability to survive by successfully modifying its

physiology and morphology to suit various hosts and can be found in various

tissues, even the nervous system. However, surprisingly little is known about the

molecular function of genes that are expressed during the course of the parasite

life cycle. To begin to decipher the molecular processes underlying gene

function, we constructed a database of expressed sequence tags (ESTs) generated

from sparganum.

FINDINGS: SpiroESTdb is a web-based information resource that is built upon the

annotation and curation of 5,655 ESTs data. SpiroESTdb provides an integrated

platform for expressed sequence data, expression dynamics, functional genes,

genetic markers including single nucleotide polymorphisms and tandem repeats,

gene ontology and KEGG pathway information. Moreover, SpiroESTdb supports easy

access to gene pages, such as (i) curation and query forms, (ii) in silico

expression profiling and (iii) BLAST search tools. Comprehensive descriptions of

the sparganum content of all sequenced data are available, including summary

reports. The contents of SpiroESTdb can be viewed and downloaded from the web

(http://pathod.cdc.go.kr/spiroestdb).

CONCLUSIONS: This integrative web-based database of sequence data, functional

annotations and expression profiling data will serve as a useful tool to help

understand and expand the characterization of parasitic infections. It can also

be used to identify potential industrial drug targets and vaccine candidate

genes.

DOI: 10.1186/1756-0500-5-130

PMCID: PMC3329409

PMID: 22397686 [Indexed for MEDLINE]

1461. BMC Bioinformatics. 2012 Mar 4;13:37. doi: 10.1186/1471-2105-13-37.

PhiSiGns: an online tool to identify signature genes in phages and design PCR

primers for examining phage diversity.

Dwivedi B(1), Schmieder R, Goldsmith DB, Edwards RA, Breitbart M.

Author information:

(1)College of Marine Science, University of South Florida, St, Petersburg, FL

33701, USA. bhaktihd@gmail.com

BACKGROUND: Phages (viruses that infect bacteria) have gained significant

attention because of their abundance, diversity and important ecological roles.

However, the lack of a universal gene shared by all phages presents a challenge

for phage identification and characterization, especially in environmental

samples where it is difficult to culture phage-host systems. Homologous conserved

genes (or "signature genes") present in groups of closely-related phages can be

used to explore phage diversity and define evolutionary relationships amongst

these phages. Bioinformatic approaches are needed to identify candidate signature

genes and design PCR primers to amplify those genes from environmental samples;

however, there is currently no existing computational tool that biologists can

use for this purpose.

RESULTS: Here we present PhiSiGns, a web-based and standalone application that

performs a pairwise comparison of each gene present in user-selected phage

genomes, identifies signature genes, generates alignments of these genes, and

designs potential PCR primer pairs. PhiSiGns is available at

(http://www.phantome.org/phisigns/; http://phisigns.sourceforge.net/) with a link

to the source code. Here we describe the specifications of PhiSiGns and

demonstrate its application with a case study.

CONCLUSIONS: PhiSiGns provides phage biologists with a user-friendly tool to

identify signature genes and design PCR primers to amplify related genes from

uncultured phages in environmental samples. This bioinformatics tool will

facilitate the development of novel signature genes for use as molecular markers

in studies of phage diversity, phylogeny, and evolution.

DOI: 10.1186/1471-2105-13-37

PMCID: PMC3314551

PMID: 22385976 [Indexed for MEDLINE]

1462. Bioinformatics. 2012 Mar 1;28(5):745-6. doi: 10.1093/bioinformatics/bts031. Epub

2012 Jan 17.

COPICAT: a software system for predicting interactions between proteins and

chemical compounds.

Sakakibara Y(1), Hachiya T, Uchida M, Nagamine N, Sugawara Y, Yokota M, Nakamura

M, Popendorf K, Komori T, Sato K.

Author information:

(1)Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Yokohama

223-8522, Japan. yasu@bio.keio.ac.jp

Since tens of millions of chemical compounds have been accumulated in public

chemical databases, fast comprehensive computational methods to predict

interactions between chemical compounds and proteins are needed for virtual

screening of lead compounds. Previously, we proposed a novel method for

predicting protein-chemical interactions using two-layer Support Vector Machine

classifiers that require only readily available biochemical data, i.e. amino acid

sequences of proteins and structure formulas of chemical compounds. In this

article, the method has been implemented as the COPICAT web service, with an

easy-to-use front-end interface. Users can simply submit a protein-chemical

interaction prediction job using a pre-trained classifier, or can even train

their own classification model by uploading training data. COPICAT's fast and

accurate computational prediction has enhanced lead compound discovery against a

database of tens of millions of chemical compounds, implying that the search

space for drug discovery is extended by >1000 times compared with currently

well-used high-throughput screening methodologies.AVAILABILITY: The COPICAT

server is available at http://copicat.dna.bio.keio.ac.jp. All functions,

including the prediction function are freely available via anonymous login

without registration. Registered users, however, can use the system more

intensively.

DOI: 10.1093/bioinformatics/bts031

PMID: 22257668 [Indexed for MEDLINE]

1463. Biotechnol Bioeng. 2012 Mar;109(3):846-50. doi: 10.1002/bit.24356. Epub 2011 Nov

9.

Compound toxicity screening and structure-activity relationship modeling in

Escherichia coli.

Planson AG(1), Carbonell P, Paillard E, Pollet N, Faulon JL.

Author information:

(1)iSSB, Institute of Systems and Synthetic Biology, University of Evry, Genopole

Campus 1, Genavenir 6, 5 rue Henri Desbruères, 91030 Evry Cedex, France.

Anne-Gaelle.Planson@issb.genopole.fr

Synthetic biology and metabolic engineering are used to develop new strategies

for producing valuable compounds ranging from therapeutics to biofuels in

engineered microorganisms. When developing methods for high-titer production

cells, toxicity is an important element to consider. Indeed the production rate

can be limited due to toxic intermediates or accumulation of byproducts of the

heterologous biosynthetic pathway of interest. Conversely, highly toxic molecules

are desired when designing antimicrobials. Compound toxicity in bacteria plays a

major role in metabolic engineering as well as in the development of new

antibacterial agents. Here, we screened a diversified chemical library of 166

compounds for toxicity in Escherichia coli. The dataset was built using a

clustering algorithm maximizing the chemical diversity in the library. The

resulting assay data was used to develop a toxicity predictor that we used to

assess the toxicity of metabolites throughout the metabolome. This new tool for

predicting toxicity can thus be used for fine-tuning heterologous expression and

can be integrated in a computational-framework for metabolic pathway design. Many

structure-activity relationship tools have been developed for toxicology studies

in eukaryotes [Valerio (2009), Toxicol Appl Pharmacol, 241(3): 356-370], however,

to the best of our knowledge we present here the first E. coli toxicity

prediction web server based on QSAR models (EcoliTox server:

http://www.issb.genopole.fr/∼faulon/EcoliTox.php).

Copyright © 2011 Wiley Periodicals, Inc.

DOI: 10.1002/bit.24356

PMID: 22038678 [Indexed for MEDLINE]

1464. Brief Bioinform. 2012 Mar;13(2):143-9. doi: 10.1093/bib/bbr044. Epub 2011 Jul 22.

LEPSCAN--a web server for searching latent periodicity in DNA sequences.

Shelenkov A(1), Korotkov E.

Author information:

(1)Bioinformatics Department, Bioengineering Centre, Russian Academy of Sciences,

Moscow, Russia. fallandar@gmail.com

A web server for searching latent periodicity based on the method of modified

profile analysis has been developed. This method allows searching latent

periodicity in presence of insertions and deletions. During searching process,

the periodicity classes are used which were found by us earlier for various

groups of organisms. Period length belongs to the range 2-20 nt, not including

the triplet periodicity. The results obtained are subjected to various filtration

steps to ensure their statistical significance.AVAILABILITY: The use of web

server is free for non-commercial users. No registration is required. URL of the

server is http://victoria.biengi.ac.ru/lepscan. Current software version is 1.06.

DOI: 10.1093/bib/bbr044

PMID: 22396486 [Indexed for MEDLINE]

1465. Brief Bioinform. 2012 Mar;13(2):175-86. doi: 10.1093/bib/bbr043. Epub 2011 Sep

10.

Dissection of human MiRNA regulatory influence to subpathway.

Li X(1), Jiang W, Li W, Lian B, Wang S, Liao M, Chen X, Wang Y, Lv Y, Wang S,

Yang L.

Author information:

(1)College of Bioinformatics Science and Technology and Bio-pharmaceutical Key

Laboratory of Heilongjiang Province, Harbin Medical University, Harbin 150081, P.

R. China. lixia@hrbmu.edu.cn

The global insight into the relationships between miRNAs and their regulatory

influences remains poorly understood. And most of complex diseases may be

attributed to certain local areas of pathway (subpathway) instead of the entire

pathway. Here, we reviewed the studies on miRNA regulations to pathways and

constructed a bipartite miRNAs and subpathways network for systematic analyzing

the miRNA regulatory influences to subpathways. We found that a small fraction of

miRNAs were global regulators, environmental information processing pathways were

preferentially regulated by miRNAs, and miRNAs had synergistic effect on

regulating group of subpathways with similar function. Integrating the disease

states of miRNAs, we also found that disease miRNAs regulated more subpathways

than nondisease miRNAs, and for all miRNAs, the number of regulated subpathways

was not in proportion to the number of the related diseases. Therefore, the study

not only provided a global view on the relationships among disease, miRNA and

subpathway, but also uncovered the function aspects of miRNA regulations and

potential pathogenesis of complex diseases. A web server to query, visualize and

download for all the data can be freely accessed at

http://bioinfo.hrbmu.edu.cn/miR2Subpath.

DOI: 10.1093/bib/bbr043

PMID: 21908864 [Indexed for MEDLINE]

1466. J Comput Aided Mol Des. 2012 Mar;26(3):339-47. doi: 10.1007/s10822-012-9560-3.

Epub 2012 Mar 18.

Real value prediction of protein folding rate change upon point mutation.

Huang LT(1), Gromiha MM.

Author information:

(1)Department of Biotechnology, Mingdao University, Changhua 523, Taiwan.

Prediction of protein folding rate change upon amino acid substitution is an

important and challenging problem in protein folding kinetics and design. In this

work, we have analyzed the relationship between amino acid properties and folding

rate change upon mutation. Our analysis showed that the correlation is not

significant with any of the studied properties in a dataset of 476 mutants.

Further, we have classified the mutants based on their locations in different

secondary structures and solvent accessibility. For each category, we have

selected a specific combination of amino acid properties using genetic algorithm

and developed a prediction scheme based on quadratic regression models for

predicting the folding rate change upon mutation. Our results showed a 10-fold

cross validation correlation of 0.72 between experimental and predicted change in

protein folding rates. The correlation is 0.73, 0.65 and 0.79, respectively in

strand, helix and coil segments. The method has been further tested with an

extended dataset of 621 mutants and a blind dataset of 62 mutants, and we

observed a good agreement with experiments. We have developed a web server for

predicting the folding rate change upon mutation and it is available at

http://bioinformatics.myweb.hinet.net/fora.htm.

DOI: 10.1007/s10822-012-9560-3

PMID: 22426539 [Indexed for MEDLINE]

1467. BMC Med Inform Decis Mak. 2012 Feb 28;12:11. doi: 10.1186/1472-6947-12-11.

Design and implementation of the first nationwide, web-based Chinese Renal Data

System (CNRDS).

Xie F(1), Zhang D, Wu J, Zhang Y, Yang Q, Sun X, Cheng J, Chen X.

Author information:

(1)Medical Systems Biology Research Center, Department of Biomedical Engineering,

Tsinghua University School of Medicine, Haidian District, Beijing 100084, China.

fengbo.xie@gmail.com

BACKGROUND: In April 2010, with an endorsement from the Ministry of Health of the

People's Republic of China, the Chinese Society of Nephrology launched the first

nationwide, web-based prospective renal data registration platform, the Chinese

Renal Data System (CNRDS), to collect structured demographic, clinical, and

laboratory data for dialysis cases, as well as to establish a kidney disease

database for researchers and policy makers.

METHODS: The CNRDS program uses information technology to facilitate healthcare

professionals to create a blood purification registry and to deliver an

evidence-based care and education protocol tailored to chronic kidney disease, as

well as online forum for communication between nephrologists. The online portal

https://www.cnrds.net is implemented as a Java web application using an Apache

Tomcat web server and a MySQL database. All data are stored in a central databank

to establish a Chinese renal database for research and publication purposes.

RESULTS: Currently, over 270,000 clinical cases, including general patient

information, diagnostics, therapies, medications, and laboratory tests, have been

registered in CNRDS by 3,669 healthcare institutions qualified for hemodialysis

therapy. At the 2011 annual blood purification forum of the Chinese Society of

Nephrology, the CNRDS 2010 annual report was reviewed and accepted by the society

members and government representatives.

CONCLUSIONS: CNRDS is the first national, web-based application for collecting

and managing electronic medical records of patients with dialysis in China. It

provides both an easily accessible platform for nephrologists to store and

organize their patient data and acts as a communication platform among

participating doctors. Moreover, it is the largest database for treatment and

patient care of end-stage renal disease (ESRD) patients in China, which will be

beneficial for scientific research and epidemiological investigations aimed at

improving the quality of life of such patients. Furthermore, it is a model

nationwide disease registry, which could potentially be used for other diseases.

DOI: 10.1186/1472-6947-12-11

PMCID: PMC3309940

PMID: 22369692 [Indexed for MEDLINE]

1468. Nat Methods. 2012 Feb 28;9(3):245-53. doi: 10.1038/nmeth.1896.

OMERO: flexible, model-driven data management for experimental biology.

Allan C(1), Burel JM, Moore J, Blackburn C, Linkert M, Loynton S, Macdonald D,

Moore WJ, Neves C, Patterson A, Porter M, Tarkowska A, Loranger B, Avondo J,

Lagerstedt I, Lianas L, Leo S, Hands K, Hay RT, Patwardhan A, Best C, Kleywegt

GJ, Zanetti G, Swedlow JR.

Author information:

(1)Wellcome Trust Centre for Gene Regulation and Expression, College of Life

Sciences, University of Dundee, Dundee, Scotland, UK.

Data-intensive research depends on tools that manage multidimensional,

heterogeneous datasets. We built OME Remote Objects (OMERO), a software platform

that enables access to and use of a wide range of biological data. OMERO uses a

server-based middleware application to provide a unified interface for images,

matrices and tables. OMERO's design and flexibility have enabled its use for

light-microscopy, high-content-screening, electron-microscopy and even

non-image-genotype data. OMERO is open-source software, available at

http://openmicroscopy.org/.

DOI: 10.1038/nmeth.1896

PMCID: PMC3437820

PMID: 22373911 [Indexed for MEDLINE]

1469. J Chem Inf Model. 2012 Feb 27;52(2):568-76. doi: 10.1021/ci2004303. Epub 2012 Feb

16.

AsteriX: a Web server to automatically extract ligand coordinates from figures in

PDF articles.

Lounnas V(1), Vriend G.

Author information:

(1)CMBI NCMLS Radboud University, Nijmegen Medical Centre, Geert Grooteplein

26-28, 6525 GA Nijmegen, The Netherlands. v-lounnas@unicancer.fr

Coordinates describing the chemical structures of small molecules that are

potential ligands for pharmaceutical targets are used at many stages of the drug

design process. The coordinates of the vast majority of ligands can be obtained

from either publicly accessible or commercial databases. However, interesting

ligands sometimes are only available from the scientific literature, in which

case their coordinates need to be reconstructed manually--a process that consists

of a series of time-consuming steps. We present a Web server that helps

reconstruct the three-dimensional (3D) coordinates of ligands for which a

two-dimensional (2D) picture is available in a PDF file. The software, called

AsteriX, analyses every picture contained in the PDF file and attempts to

determine automatically whether or not it contains ligands. Areas in pictures

that may contain molecular structures are processed to extract connectivity and

atom type information that allow coordinates to be subsequently reconstructed.

The AsteriX Web server was tested on a series of articles containing a large

diversity in graphical representations. In total, 88% of 3249 ligand structures

present in the test set were identified as chemical diagrams. Of these, about

half were interpreted correctly as 3D structures, and a further one-third

required only minor manual corrections. It is principally impossible to always

correctly reconstruct 3D coordinates from pictures because there are many

different protocols for drawing a 2D image of a ligand, but more importantly a

wide variety of semantic annotations are possible. The AsteriX Web server

therefore includes facilities that allow the users to augment partial or

partially correct 3D reconstructions. All 3D reconstructions are submitted,

checked, and corrected by the users domain at the server and are freely available

for everybody. The coordinates of the reconstructed ligands are made available in

a series of formats commonly used in drug design research. The AsteriX Web server

is freely available at http://swift.cmbi.ru.nl/bitmapb/.

DOI: 10.1021/ci2004303

PMID: 22299625 [Indexed for MEDLINE]

1470. Bioinformatics. 2012 Feb 15;28(4):516-22. doi: 10.1093/bioinformatics/btr710.

Epub 2012 Jan 13.

BOCTOPUS: improved topology prediction of transmembrane β barrel proteins.

Hayat S(1), Elofsson A.

Author information:

(1)Center for Biomembrane Research, Department of Biochemistry and Biophysics,

Stockholm Bioinformatics Center, SciLifeLab, Swedish E-science Research Center,

Stockholm University, SE-10691 Stockholm, Sweden.

MOTIVATION: Transmembrane β barrel proteins (TMBs) are found in the outer

membrane of Gram-negative bacteria, chloroplast and mitochondria. They play a

major role in the translocation machinery, pore formation, membrane anchoring and

ion exchange. TMBs are also promising targets for antimicrobial drugs and

vaccines. Given the difficulty in membrane protein structure determination,

computational methods to identify TMBs and predict the topology of TMBs are

important.

RESULTS: Here, we present BOCTOPUS; an improved method for the topology

prediction of TMBs by employing a combination of support vector machines (SVMs)

and Hidden Markov Models (HMMs). The SVMs and HMMs account for local and global

residue preferences, respectively. Based on a 10-fold cross-validation test,

BOCTOPUS performs better than all existing methods, reaching a Q3 accuracy of

87%. Further, BOCTOPUS predicted the correct number of strands for 83% proteins

in the dataset. BOCTOPUS might also help in reliable identification of TMBs by

using it as an additional filter to methods specialized in this task.

AVAILABILITY: BOCTOPUS is freely available as a web server at:

http://boctopus.cbr.su.se/. The datasets used for training and evaluations are

also available from this site.

DOI: 10.1093/bioinformatics/btr710

PMID: 22247276 [Indexed for MEDLINE]

1471. Bioinformatics. 2012 Feb 15;28(4):591-2. doi: 10.1093/bioinformatics/btr706. Epub

2011 Dec 22.

LaTcOm: a web server for visualizing rare codon clusters in coding sequences.

Theodosiou A(1), Promponas VJ.

Author information:

(1)Bioinformatics Research Laboratory, Department of Biological Sciences,

University of Cyprus, PO Box 20537, CY 1678, Nicosia, Cyprus.

We present LaTcOm, a new web tool, which offers several alternative methods for

'rare codon cluster' (RCC) identification from a single and simple graphical user

interface. In the current version, three RCC detection schemes are implemented:

the recently described %MinMax algorithm and a simplified sliding window

approach, along with a novel modification of a linear-time algorithm for the

detection of maximally scoring subsequences tailored to the RCC detection

problem. Among a number of user tunable parameters, several codon-based scales

relevant for RCC detection are available, including tRNA abundance values from

Escherichia coli and several codon usage tables from a selection of genomes.

Furthermore, useful scale transformations may be performed upon user request

(e.g. linear, sigmoid). Users may choose to visualize RCC positions within the

submitted sequences either with graphical representations or in textual form for

further processing.AVAILABILITY: LaTcOm is freely available online at the URL

http://troodos.biol.ucy.ac.cy/latcom.html.

DOI: 10.1093/bioinformatics/btr706

PMID: 22199385 [Indexed for MEDLINE]

1472. Bioinformatics. 2012 Feb 15;28(4):503-9. doi: 10.1093/bioinformatics/btr682. Epub

2011 Dec 20.

ESpritz: accurate and fast prediction of protein disorder.

Walsh I(1), Martin AJ, Di Domenico T, Tosatto SC.

Author information:

(1)Department of Biology, University of Padua, Viale G. Colombo 3, I-35131

Padova, Italy.

MOTIVATION: Intrinsically disordered regions are key for the function of numerous

proteins, and the scant available experimental annotations suggest the existence

of different disorder flavors. While efficient predictions are required to

annotate entire genomes, most existing methods require sequence profiles for

disorder prediction, making them cumbersome for high-throughput applications.

RESULTS: In this work, we present an ensemble of protein disorder predictors

called ESpritz. These are based on bidirectional recursive neural networks and

trained on three different flavors of disorder, including a novel NMR flexibility

predictor. ESpritz can produce fast and accurate sequence-only predictions,

annotating entire genomes in the order of hours on a single processor core.

Alternatively, a slower but slightly more accurate ESpritz variant using sequence

profiles can be used for applications requiring maximum performance. Two levels

of prediction confidence allow either to maximize reasonable disorder detection

or to limit expected false positives to 5%. ESpritz performs consistently well on

the recent CASP9 data, reaching a S(w) measure of 54.82 and area under the

receiver operator curve of 0.856. The fast predictor is four orders of magnitude

faster and remains better than most publicly available CASP9 methods, making it

ideal for genomic scale predictions.

CONCLUSIONS: ESpritz predicts three flavors of disorder at two distinct false

positive rates, either with a fast or slower and slightly more accurate approach.

Given its state-of-the-art performance, it can be especially useful for

high-throughput applications.

AVAILABILITY: Both a web server for high-throughput analysis and a Linux

executable version of ESpritz are available from:

http://protein.bio.unipd.it/espritz/.

DOI: 10.1093/bioinformatics/btr682

PMID: 22190692 [Indexed for MEDLINE]

1473. BMC Bioinformatics. 2012 Feb 15;13:33. doi: 10.1186/1471-2105-13-33.

BLANNOTATOR: enhanced homology-based function prediction of bacterial proteins.

Kankainen M(1), Ojala T, Holm L.

Author information:

(1)Institute of Biotechnology, University of Helsinki, Helsinki, Finland.

matti.kankainen@helsinki.fi

BACKGROUND: Automated function prediction has played a central role in

determining the biological functions of bacterial proteins. Typically, protein

function annotation relies on homology, and function is inferred from other

proteins with similar sequences. This approach has become popular in bacterial

genomics because it is one of the few methods that is practical for large

datasets and because it does not require additional functional genomics

experiments. However, the existing solutions produce erroneous predictions in

many cases, especially when query sequences have low levels of identity with the

annotated source protein. This problem has created a pressing need for

improvements in homology-based annotation.

RESULTS: We present an automated method for the functional annotation of

bacterial protein sequences. Based on sequence similarity searches, BLANNOTATOR

accurately annotates query sequences with one-line summary descriptions of

protein function. It groups sequences identified by BLAST into subsets according

to their annotation and bases its prediction on a set of sequences with

consistent functional information. We show the results of BLANNOTATOR's

performance in sets of bacterial proteins with known functions. We simulated the

annotation process for 3090 SWISS-PROT proteins using a database in its state

preceding the functional characterisation of the query protein. For this dataset,

our method outperformed the five others that we tested, and the improved

performance was maintained even in the absence of highly related sequence hits.

We further demonstrate the value of our tool by analysing the putative proteome

of Lactobacillus crispatus strain ST1.

CONCLUSIONS: BLANNOTATOR is an accurate method for bacterial protein function

prediction. It is practical for genome-scale data and does not require

pre-existing sequence clustering; thus, this method suits the needs of bacterial

genome and metagenome researchers. The method and a web-server are available at

http://ekhidna.biocenter.helsinki.fi/poxo/blannotator/.

DOI: 10.1186/1471-2105-13-33

PMCID: PMC3386020

PMID: 22335941 [Indexed for MEDLINE]

1474. BMC Res Notes. 2012 Feb 13;5:92. doi: 10.1186/1756-0500-5-92.

miRviewer: a multispecies microRNA homologous viewer.

Kiezun A(1), Artzi S, Modai S, Volk N, Isakov O, Shomron N.

Author information:

(1)Department of Cell and Developmental Biology, Sackler Faculty of Medicine, Tel

Aviv University, Tel Aviv 69978, Israel. nshomron@post.tau.ac.il.

BACKGROUND: MicroRNAs (miRNAs) are short non-coding RNAs that regulate gene

expression via binding to the 3' ends of mRNAs. MiRNAs have been associated with

many cellular events ascertaining their central role in gene regulation. In order

to better understand miRNAs of interest it is of utmost importance to learn about

the genomic conservation of these genes.

FINDINGS: The miRviewer web-server, presented here, encompasses all known miRNAs

of currently fully annotated animal genomes in a visual 'birds-eye' view

representation. miRviewer provides a graphical outlook of the current miRNA world

together with sequence alignments and secondary structures of each miRNA. As a

test case we experimentally examined the expression of several miRNAs in various

animals.

CONCLUSIONS: miRviewer completes the homologous miRNA space with hundreds of

unreported miRNAs and is available at:

http://people.csail.mit.edu/akiezun/miRviewer.

DOI: 10.1186/1756-0500-5-92

PMCID: PMC3292992

PMID: 22330228

1475. J Clin Oncol. 2012 Feb 10;30(5):497-506. doi: 10.1200/JCO.2011.38.6060. Epub 2012

Jan 9.

Online tool to guide decisions for BRCA1/2 mutation carriers.

Kurian AW(1), Munoz DF, Rust P, Schackmann EA, Smith M, Clarke L, Mills MA,

Plevritis SK.

Author information:

(1)Stanford University School of Medicine, Stanford, CA, USA.

Comment in

J Clin Oncol. 2012 Feb 10;30(5):471-3.

PURPOSE: Women with BRCA1 or BRCA2 (BRCA1/2) mutations must choose between

prophylactic surgeries and screening to manage their high risks of breast and

ovarian cancer, comparing options in terms of cancer incidence, survival, and

quality of life. A clinical decision tool could guide these complex choices.

METHODS: We built a Monte Carlo model for BRCA1/2 mutation carriers, simulating

breast screening with annual mammography plus magnetic resonance imaging (MRI)

from ages 25 to 69 years and prophylactic mastectomy (PM) and/or prophylactic

oophorectomy (PO) at various ages. Modeled outcomes were cancer incidence, tumor

features that shape treatment recommendations, overall survival, and

cause-specific mortality. We adapted the model into an online tool to support

shared decision making.

RESULTS: We compared strategies on cancer incidence and survival to age 70 years;

for example, PO plus PM at age 25 years optimizes both outcomes (incidence, 4% to

11%; survival, 80% to 83%), whereas PO at age 40 years plus MRI screening offers

less effective prevention, yet similar survival (incidence, 36% to 57%; survival,

74% to 80%). To characterize patients' treatment and survivorship experiences, we

reported the tumor features and treatments associated with risk-reducing

interventions; for example, in most BRCA2 mutation carriers (81%), MRI screening

diagnoses stage I, hormone receptor-positive breast cancers, which may not

require chemotherapy.

CONCLUSION: Cancer risk-reducing options for BRCA1/2 mutation carriers vary in

their impact on cancer incidence, recommended treatments, quality of life, and

survival. To guide decisions informed by multiple health outcomes, we provide an

online tool for joint use by patients with their physicians

(http://brcatool.stanford.edu).

DOI: 10.1200/JCO.2011.38.6060

PMCID: PMC3295552

PMID: 22231042 [Indexed for MEDLINE]

1476. Bioinformatics. 2012 Feb 1;28(3):306-10. doi: 10.1093/bioinformatics/btr672. Epub

2011 Dec 6.

InFiRe -- a novel computational method for the identification of insertion sites

in transposon mutagenized bacterial genomes.

Shevchuk O(1), Roselius L, Günther G, Klein J, Jahn D, Steinert M, Münch R.

Author information:

(1)Institute of Microbiology, Technische Universität Braunschweig,

Spielmannstrasse 7, 38106 Braunschweig, Germany.

MOTIVATION: InFiRe, Insertion Finder via Restriction digest, is a novel software

tool that allows for the computational identification of transposon insertion

sites in known bacterial genome sequences after transposon mutagenesis

experiments. The approach is based on the fact that restriction endonuclease

digestions of bacterial DNA yield a unique pattern of DNA fragments with defined

sizes. Transposon insertion changes the size of the hosting DNA fragment by a

known number of base pairs. The exact size of this fragment can be determined by

Southern blot hybridization. Subsequently, the position of insertion can be

identified with computational analysis. The outlined method provides a solid

basis for the establishment of a new high-throughput technology.

AVAILABILITY AND IMPLEMENTATION: The software is freely available on our web

server at www.infire.tu-bs.de. The algorithm was implemented in the statistical

programming language R. For the most flexible use, InFiRe is provided in two

different versions. A web interface offers the convenient use in a web browser.

In addition, the software and source code is freely available for download as

R-packages on our website.

CONTACT: m.steinert@tu-bs.de

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr672

PMID: 22155866 [Indexed for MEDLINE]

1477. Bioinformatics. 2012 Feb 1;28(3):448-50. doi: 10.1093/bioinformatics/btr662. Epub

2011 Nov 29.

An infrastructure for ontology-based information systems in biomedicine: RICORDO

case study.

Wimalaratne SM(1), Grenon P, Hoehndorf R, Gkoutos GV, de Bono B.

Author information:

(1)European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge

CB10 1SD, UK. sarala@ebi.ac.uk

SUMMARY: The article presents an infrastructure for supporting the semantic

interoperability of biomedical resources based on the management (storing and

inference-based querying) of their ontology-based annotations. This

infrastructure consists of: (i) a repository to store and query ontology-based

annotations; (ii) a knowledge base server with an inference engine to support the

storage of and reasoning over ontologies used in the annotation of resources;

(iii) a set of applications and services allowing interaction with the integrated

repository and knowledge base. The infrastructure is being prototyped and

developed and evaluated by the RICORDO project in support of the knowledge

management of biomedical resources, including physiology and pharmacology models

and associated clinical data.

AVAILABILITY AND IMPLEMENTATION: The RICORDO toolkit and its source code are

freely available from http://ricordo.eu/relevant-resources.

CONTACT: sarala@ebi.ac.uk.

DOI: 10.1093/bioinformatics/btr662

PMID: 22130590 [Indexed for MEDLINE]

1478. Comput Biol Med. 2012 Feb;42(2):228-34. doi: 10.1016/j.compbiomed.2011.11.012.

Epub 2011 Dec 20.

THEME: a web tool for loop-design microarray data analysis.

Chen CR(1), Shu WY, Tsai ML, Cheng WC, Hsu IC.

Author information:

(1)Department of Biomedical Engineering and Environmental Sciences, National

Tsing Hua University, Hsinchu, Taiwan. d948504@oz.nthu.edu.tw

A number of recent studies have shown that loop-design is more efficient than

reference control design. Data analysis for loop-design microarray experiments is

commonly undertaken using linear models and statistical tests. These techniques

require specialized knowledge in statistical programming. However, limited

loop-design web-based tools are available. We have developed the THEME (Tsing Hua

Engine of Microarray Experiment) that exploits all necessary data analysis tools

for loop-design microarray studies. THEME allows users to construct linear models

and to apply multiple user-defined statistical tests of hypotheses for detection

of DEG (differentially expressed genes). Users can modify entries of design

matrix for experimental design as well as that of contrast matrix for statistical

tests of hypotheses. The output of multiple user-defined statistical tests of

hypotheses, DEG lists, can be cross-validated. The web platform provides data

assessment and visualization tools that significantly assist users when

evaluating the performance of microarray experimental procedures. THEME is also a

MIAME (Minimal Information About a Microarray Experiment) compliant system, which

enables users to export formatted files for GEO (Gene Expression Omnibus)

submission. THEME offers comprehensive web services to biologists for data

analysis of loop-design microarray experiments. This web-based resource is

especially useful for core facility service as well as collaboration projects

when researchers are not at the same site. Data analysis procedures, starting

from uploading raw data files to retrieving DEG lists, can be flexibly operated

with natural workflows. These features make THEME a reliable and powerful on-line

system for data analysis of loop-design microarrays. The THEME server is

available at http://metadb.bmes.nthu.edu.tw/theme/.

Copyright © 2011 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiomed.2011.11.012

PMID: 22188791 [Indexed for MEDLINE]

1479. Fungal Genet Biol. 2012 Feb;49(2):173-9. doi: 10.1016/j.fgb.2011.12.009. Epub

2012 Jan 3.

ProFASTA: a pipeline web server for fungal protein scanning with integration of

cell surface prediction software.

de Groot PW(1), Brandt BW.

Author information:

(1)Regional Center for Biomedical Research, Albacete Science and Technology Park,

University of Castilla - La Mancha, 02006 Albacete, Spain. Piet.deGroot@uclm.es

Surface proteins, such as those located in the cell wall of fungi, play an

important role in the interaction with the surrounding environment. For instance,

they mediate primary host-pathogen interactions and are crucial to the

establishment of biofilms and fungal infections. Surface localization of proteins

is determined by specific sequence features and can be predicted by combining

different freely available web servers. However, user-friendly tools that allow

rapid analysis of large datasets (whole proteomes or larger) in subsequent

analyses were not yet available. Here, we present the web tool ProFASTA, which

integrates multiple tools for rapid scanning of protein sequence properties in

large datasets and returns sequences in FASTA format. ProFASTA also allows for

pipeline filtering of proteins with cell surface characteristics by analysis of

the output created with SignalP, TMHMM and big-PI. In addition, it provides

keyword, iso-electric point, composition and pattern scanning. Furthermore,

ProFASTA contains all fungal protein sequences present in the NCBI Protein

database. As the full fungal NCBI Taxonomy is included, sequence subsets can be

selected by supplying a taxon name. The usefulness of ProFASTA is demonstrated

here with a few examples; in the recent past, ProFASTA has already been applied

successfully to the annotation of covalently-bound fungal wall proteins as part

of community-wide genome annotation programs. ProFASTA is available at:

http://www.bioinformatics.nl/tools/profasta/.

Copyright © 2012 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.fgb.2011.12.009

PMID: 22230096 [Indexed for MEDLINE]

1480. Hum Mutat. 2012 Feb;33(2):332-7. doi: 10.1002/humu.21642. Epub 2011 Dec 9.

Hansa: an automated method for discriminating disease and neutral human nsSNPs.

Acharya V(1), Nagarajaram HA.

Author information:

(1)Laboratory of Computational Biology, Centre for DNA Fingerprinting &

Diagnostics, Bldg7, Gruhakalpa, Nampally, Hyderabad, 500001, Andhra Pradesh,

India.

Comment in

Hum Mutat. 2013 Feb;34(2):405-6.

Hum Mutat. 2013 Feb;34(2):407.

Variations are mostly due to nonsynonymous single nucleotide polymorphisms

(nsSNPs), some of which are associated with certain diseases. Phenotypic effects

of a large number of nsSNPs have not been characterized. Although several methods

have been developed to predict the effects of nsSNPs as "disease" or "neutral,"

there is still a need for development of methods with improved prediction

accuracies. We, therefore, developed a support vector machine (SVM) based method

named Hansa which uses a novel set of discriminatory features to classify nsSNPs

into disease (pathogenic) and benign (neutral) types. Validation studies on a

benchmark dataset and further on an independent dataset of well-characterized

known disease and neutral mutations show that Hansa outperforms the other known

methods. For example, fivefold cross-validation studies using the benchmark

HumVar dataset reveal that at the false positive rate (FPR) of 20% Hansa yields a

true positive rate (TPR) of 82% that is about 10% higher than the best-known

method. Hansa is available in the form of a web server at

http://hansa.cdfd.org.in:8080.

© 2011 Wiley Periodicals, Inc.

DOI: 10.1002/humu.21642

PMID: 22045683 [Indexed for MEDLINE]

1481. J Bioinform Comput Biol. 2012 Feb;10(1):1240005. doi: 10.1142/S0219720012400057.

Generation of synthetic data and experimental designs in evaluating interactions

for association studies.

Chen Y(1), Li J.

Author information:

(1)Department of Electrical Engineering and Computer Science, Case Western

Reserve University, 10900 Euclid Ave, Cleveland, OH 44106, USA.

yixuan.chen@case.edu

Complex diseases, by definition, involve multiple factors, including gene-gene

interactions and gene-environment interactions. Researchers commonly rely on

simulated data to evaluate their approaches for detecting high-order interactions

in disease gene mapping. A publicly available simulation program to generate

samples involving complex genetic and environmental interactions is of great

interest to the community. We have developed a software package named gs1.0,

which has been widely used since its publication. In this article, we present an

upgraded version gs2.0, which not only inherits its capacity to generate

realistic genotype data but also provides great functionality and flexibility to

simulate various interaction models. In addition to a standalone version, a

user-friendly web server (http://cbc.case.edu/gs) has been set up to help users

to build complex interaction models. Furthermore, by utilizing three three-locus

models as an example, we have shown how realistic model parameters can be chosen

in generating simulated data.

DOI: 10.1142/S0219720012400057

PMID: 22809306 [Indexed for MEDLINE]

1482. Proteins. 2012 Feb;80(2):352-61. doi: 10.1002/prot.23183. Epub 2011 Nov 22.

Template-based protein structure modeling using TASSER(VMT.).

Zhou H(1), Skolnick J.

Author information:

(1)Center for the Study of Systems Biology, School of Biology, Georgia Institute

of Technology, Atlanta, Georgia 30318.

Template-based protein structure modeling is commonly used for protein structure

prediction. Based on the observation that multiple template-based methods often

perform better than single template-based methods, we further explore the use of

a variable number of multiple templates for a given target in the latest variant

of TASSER, TASSER(VMT) . We first develop an algorithm that improves the

target-template alignment for a given template. The improved alignment, called

the SP(3) alternative alignment, is generated by a parametric alignment method

coupled with short TASSER refinement on models selected using knowledge-based

scores. The refined top model is then structurally aligned to the template to

produce the SP(3) alternative alignment. Templates identified using SP(3)

threading are combined with the SP(3) alternative and HHEARCH alignments to

provide target alignments to each template. These template models are then

grouped into sets containing a variable number of template/alignment

combinations. For each set, we run short TASSER simulations to build full-length

models. Then, the models from all sets of templates are pooled, and the top 20-50

models selected using FTCOM ranking method. These models are then subjected to a

single longer TASSER refinement run for final prediction. We benchmarked our

method by comparison with our previously developed approach, pro-sp(3) -TASSER,

on a set with 874 easy and 318 hard targets. The average GDT-TS score

improvements for the first model are 3.5 and 4.3% for easy and hard targets,

respectively. When tested on the 112 CASP9 targets, our method improves the

average GDT-TS scores as compared to pro-sp3-TASSER by 8.2 and 9.3% for the 80

easy and 32 hard targets, respectively. It also shows slightly better results

than the top ranked CASP9 Zhang-Server, QUARK and HHpredA methods. The program is

available for download at http://cssb.biology.gatech.edu/.

Copyright © 2011 Wiley Periodicals, Inc.

DOI: 10.1002/prot.23183

PMCID: PMC3291807

PMID: 22105797 [Indexed for MEDLINE]

1483. Proteins. 2012 Feb;80(2):374-81. doi: 10.1002/prot.23188. Epub 2011 Nov 17.

Prediction of protein secondary structure from circular dichroism using

theoretically derived spectra.

Louis-Jeune C(1), Andrade-Navarro MA, Perez-Iratxeta C.

Author information:

(1)Sprott Center for Stem Cell Research, Ottawa Hospital Research Institute,

Ottawa, Ontario K1H 8L6, Canada.

Erratum in

Proteins. 2012 Dec;80(12):2818.

Circular dichroism (CD) is a spectroscopic technique commonly used to investigate

the structure of proteins. Major secondary structure types, alpha-helices and

beta-strands, produce distinctive CD spectra. Thus, by comparing the CD spectrum

of a protein of interest to a reference set consisting of CD spectra of proteins

of known structure, predictive methods can estimate the secondary structure of

the protein. Currently available methods, including K2D2, use such experimental

CD reference sets, which are very small in size when compared to the number of

tertiary structures available in the Protein Data Bank (PDB). Conversely, given a

PDB structure, it is possible to predict a theoretical CD spectrum from it. The

methodological framework for this calculation was established long ago but only

recently a convenient implementation called DichroCalc has been developed. In

this study, we set to determine whether theoretically derived spectra could be

used as reference set for accurate CD based predictions of secondary structure.

We used DichroCalc to calculate the theoretical CD spectra of a nonredundant set

of structures representing most proteins in the PDB, and applied a

straightforward approach for predicting protein secondary structure content using

these theoretical CD spectra as reference set. We show that this method improves

the predictions, particularly for the wavelength interval between 200 and 240 nm

and for beta-strand content. We have implemented this method, called K2D3, in a

publicly accessible web server at http://www. ogic.ca/projects/k2d3.

Copyright © 2011 Wiley Periodicals, Inc.

DOI: 10.1002/prot.23188

PMID: 22095872 [Indexed for MEDLINE]

1484. J Comput Chem. 2012 Jan 30;33(3):259-67. doi: 10.1002/jcc.21968. Epub 2011 Nov 2.

SPINE X: improving protein secondary structure prediction by multistep learning

coupled with prediction of solvent accessible surface area and backbone torsion

angles.

Faraggi E(1), Zhang T, Yang Y, Kurgan L, Zhou Y.

Author information:

(1)School of Informatics, Indiana University Purdue University, Indianapolis,

Indiana, USA.

Accurate prediction of protein secondary structure is essential for accurate

sequence alignment, three-dimensional structure modeling, and function

prediction. The accuracy of ab initio secondary structure prediction from

sequence, however, has only increased from around 77 to 80% over the past decade.

Here, we developed a multistep neural-network algorithm by coupling secondary

structure prediction with prediction of solvent accessibility and backbone

torsion angles in an iterative manner. Our method called SPINE X was applied to a

dataset of 2640 proteins (25% sequence identity cutoff) previously built for the

first version of SPINE and achieved a 82.0% accuracy based on 10-fold cross

validation (Q(3)). Surpassing 81% accuracy by SPINE X is further confirmed by

employing an independently built test dataset of 1833 protein chains, a recently

built dataset of 1975 proteins and 117 CASP 9 targets (critical assessment of

structure prediction techniques) with an accuracy of 81.3%, 82.3% and 81.8%,

respectively. The prediction accuracy is further improved to 83.8% for the

dataset of 2640 proteins if the DSSP assignment used above is replaced by a more

consistent consensus secondary structure assignment method. Comparison to the

popular PSIPRED and CASP-winning structure-prediction techniques is made. SPINE X

predicts number of helices and sheets correctly for 21.0% of 1833 proteins,

compared to 17.6% by PSIPRED. It further shows that SPINE X consistently makes

more accurate prediction in helical residues (6%) without over prediction while

PSIPRED makes more accurate prediction in coil residues (3-5%) and over predicts

them by 7%. SPINE X Server and its training/test datasets are available at

http://sparks.informatics.iupui.edu/

Copyright © 2011 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.21968

PMCID: PMC3240697

PMID: 22045506 [Indexed for MEDLINE]

1485. Genet Mol Res. 2012 Jan 27;11(1):174-81. doi: 10.4238/2012.January.27.4.

Improved method for predicting protein fold patterns with ensemble classifiers.

Chen W(1), Liu X, Huang Y, Jiang Y, Zou Q, Lin C.

Author information:

(1)School of Information Science and Technology, Xiamen University, Xiamen,

Fujian, China.

Protein folding is recognized as a critical problem in the field of biophysics in

the 21st century. Predicting protein-folding patterns is challenging due to the

complex structure of proteins. In an attempt to solve this problem, we employed

ensemble classifiers to improve prediction accuracy. In our experiments,

188-dimensional features were extracted based on the composition and

physical-chemical property of proteins and 20-dimensional features were selected

using a coupled position-specific scoring matrix. Compared with traditional

prediction methods, these methods were superior in terms of prediction accuracy.

The 188-dimensional feature-based method achieved 71.2% accuracy in five

cross-validations. The accuracy rose to 77% when we used a 20-dimensional feature

vector. These methods were used on recent data, with 54.2% accuracy. Source codes

and dataset, together with web server and software tools for prediction, are

available at: http://datamining.xmu.edu.cn/main/~cwc/ProteinPredict.html.

DOI: 10.4238/2012.January.27.4

PMID: 22370884 [Indexed for MEDLINE]

1486. Bioinformatics. 2012 Jan 15;28(2):286-7. doi: 10.1093/bioinformatics/btr651. Epub

2011 Nov 22.

FTSite: high accuracy detection of ligand binding sites on unbound protein

structures.

Ngan CH(1), Hall DR, Zerbe B, Grove LE, Kozakov D, Vajda S.

Author information:

(1)Department of Biomedical Engineering, Boston University, Boston, MA 02115,

USA.

MOTIVATION: Binding site identification is a classical problem that is important

for a range of applications, including the structure-based prediction of

function, the elucidation of functional relationships among proteins, protein

engineering and drug design. We describe an accurate method of binding site

identification, namely FTSite. This method is based on experimental evidence that

ligand binding sites also bind small organic molecules of various shapes and

polarity. The FTSite algorithm does not rely on any evolutionary or statistical

information, but achieves near experimental accuracy: it is capable of

identifying the binding sites in over 94% of apo proteins from established test

sets that have been used to evaluate many other binding site prediction methods.

AVAILABILITY: FTSite is freely available as a web-based server at

http://ftsite.bu.edu.

DOI: 10.1093/bioinformatics/btr651

PMCID: PMC3259439

PMID: 22113084 [Indexed for MEDLINE]

1487. Bioinformatics. 2012 Jan 15;28(2):198-205. doi: 10.1093/bioinformatics/btr636.

Epub 2011 Nov 21.

MetalionRNA: computational predictor of metal-binding sites in RNA structures.

Philips A(1), Milanowska K, Lach G, Boniecki M, Rother K, Bujnicki JM.

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland.

MOTIVATION: Metal ions are essential for the folding of RNA molecules into stable

tertiary structures and are often involved in the catalytic activity of

ribozymes. However, the positions of metal ions in RNA 3D structures are

difficult to determine experimentally. This motivated us to develop a

computational predictor of metal ion sites for RNA structures.

RESULTS: We developed a statistical potential for predicting positions of metal

ions (magnesium, sodium and potassium), based on the analysis of binding sites in

experimentally solved RNA structures. The MetalionRNA program is available as a

web server that predicts metal ions for RNA structures submitted by the user.

AVAILABILITY: The MetalionRNA web server is accessible at

http://metalionrna.genesilico.pl/.

DOI: 10.1093/bioinformatics/btr636

PMCID: PMC3259437

PMID: 22110243 [Indexed for MEDLINE]

1488. J Theor Biol. 2012 Jan 7;292:93-102. doi: 10.1016/j.jtbi.2011.09.026. Epub 2011

Oct 6.

MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM.

Hayat M(1), Khan A.

Author information:

(1)Department of Computer and Information Sciences, Pakistan Institute of

Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan.

About 50% of available drugs are targeted against membrane proteins. Knowledge of

membrane protein's structure and function has great importance in biological and

pharmacological research. Therefore, an automated method is exceedingly

advantageous, which can help in identifying the new membrane protein types based

on their primary sequence. In this paper, we tackle the interesting problem of

classifying membrane protein types using their sequence information. We consider

both evolutionary and physicochemical features and provide them to our

classification system based on support vector machine (SVM) with error correction

code. We employ a powerful sequence encoding scheme by fusing position specific

scoring matrix and split amino acid composition to effectively discriminate

membrane protein types. Linear, polynomial, and RBF based-SVM with Bose,

Chaudhuri, Hocquenghem coding are trained and tested. The highest success rate of

91.1% and 93.4% on two datasets is obtained by RBF-SVM using leave-one-out

cross-validation. Thus, our proposed approach is an effective tool for the

discrimination of membrane protein types and might be helpful to

researchers/academicians working in the field of Drug Discovery, Cell Biology,

and Bioinformatics. The web server for the proposed MemHyb-SVM is accessible at

http://111.68.99.218/MemHyb-SVM.

Copyright © 2011 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2011.09.026

PMID: 22001079 [Indexed for MEDLINE]

1489. Adv Appl Bioinform Chem. 2012;5:11-21. doi: 10.2147/AABC.S30620. Epub 2012 Jul

25.

B-Pred, a structure based B-cell epitopes prediction server.

Giacò L(1), Amicosante M, Fraziano M, Gherardini PF, Ausiello G, Helmer-Citterich

M, Colizzi V, Cabibbo A.

Author information:

(1)Department of Biology, University of Rome "Tor Vergata", Rome, Italy.

The ability to predict immunogenic regions in selected proteins by in-silico

methods has broad implications, such as allowing a quick selection of potential

reagents to be used as diagnostics, vaccines, immunotherapeutics, or research

tools in several branches of biological and biotechnological research. However,

the prediction of antibody target sites in proteins using computational

methodologies has proven to be a highly challenging task, which is likely due to

the somewhat elusive nature of B-cell epitopes. This paper proposes a web-based

platform for scoring potential immunological reagents based on the structures or

3D models of the proteins of interest. The method scores a protein's peptides

set, which is derived from a sliding window, based on the average solvent

exposure, with a filter on the average local model quality for each peptide. The

platform was validated on a custom-assembled database of 1336 experimentally

determined epitopes from 106 proteins for which a reliable 3D model could be

obtained through standard modeling techniques. Despite showing poor sensitivity,

this method can achieve a specificity of 0.70 and a positive predictive value of

0.29 by combining these two simple parameters. These values are slightly higher

than those obtained with other established sequence-based or structure-based

methods that have been evaluated using the same epitopes dataset. This method is

implemented in a web server called B-Pred, which is accessible at

http://immuno.bio.uniroma2.it/bpred. The server contains a number of original

features that allow users to perform personalized reagent searches by

manipulating the sliding window's width and sliding step, changing the exposure

and model quality thresholds, and running sequential queries with different

parameters. The B-Pred server should assist experimentalists in the rational

selection of epitope antigens for a wide range of applications.

DOI: 10.2147/AABC.S30620

PMCID: PMC3413014

PMID: 22888263

1490. Am J Physiol Cell Physiol. 2012 Jan 1;302(1):C154-64. doi:

10.1152/ajpcell.00325.2011. Epub 2011 Sep 28.

NHLBI-AbDesigner: an online tool for design of peptide-directed antibodies.

Pisitkun T(1), Hoffert JD, Saeed F, Knepper MA.

Author information:

(1)Epithelial Systems Biology Laboratory, National Heart, Lung, and Blood

Institute, National Institutes of Health, Bethesda, Maryland 20892-1603, USA.

Investigation of physiological mechanisms at a cellular level often requires

production of high-quality antibodies, frequently using synthetic peptides as

immunogens. Here we describe a new, web-based software tool called

NHLBI-AbDesigner that allows the user to visualize the information needed to

choose optimal peptide sequences for peptide-directed antibody production

(http://helixweb.nih.gov/AbDesigner/). The choice of an immunizing peptide is

generally based on a need to optimize immunogenicity, antibody specificity,

multispecies conservation, and robustness in the face of posttranslational

modifications (PTMs). AbDesigner displays information relevant to these criteria

as follows: 1) "Immunogenicity Score," based on hydropathy and secondary

structure prediction; 2) "Uniqueness Score," a predictor of specificity of an

antibody against all proteins expressed in the same species; 3) "Conservation

Score," a predictor of ability of the antibody to recognize orthologs in other

animal species; and 4) "Protein Features" that show structural domains, variable

regions, and annotated PTMs that may affect antibody performance. AbDesigner

displays the information online in an interactive graphical user interface, which

allows the user to recognize the trade-offs that exist for alternative synthetic

peptide choices and to choose the one that is best for a proposed application.

Several examples of the use of AbDesigner for the display of such trade-offs are

presented, including production of a new antibody to Slc9a3. We also used the

program in large-scale mode to create a database listing the 15-amino acid

peptides with the highest Immunogenicity Scores for all known proteins in five

animal species, one plant species (Arabidopsis thaliana), and Saccharomyces

cerevisiae.

DOI: 10.1152/ajpcell.00325.2011

PMCID: PMC3328907

PMID: 21956165 [Indexed for MEDLINE]

1491. AMIA Annu Symp Proc. 2012;2012:1350-9. Epub 2012 Nov 3.

A collaborative framework for Distributed Privacy-Preserving Support Vector

Machine learning.

Que J(1), Jiang X, Ohno-Machado L.

Author information:

(1)University of California, La Jolla, CA, USA.

A Support Vector Machine (SVM) is a popular tool for decision support. The

traditional way to build an SVM model is to estimate parameters based on a

centralized repository of data. However, in the field of biomedicine, patient

data are sometimes stored in local repositories or institutions where they were

collected, and may not be easily shared due to privacy concerns. This creates a

substantial barrier for researchers to effectively learn from the distributed

data using machine learning tools like SVMs. To overcome this difficulty and

promote efficient information exchange without sharing sensitive raw data, we

developed a Distributed Privacy Preserving Support Vector Machine (DPP-SVM). The

DPP-SVM enables privacy-preserving collaborative learning, in which a trusted

server integrates "privacy-insensitive" intermediary results. The globally

learned model is guaranteed to be exactly the same as learned from combined data.

We also provide a free web-service (http://privacy.ucsd.edu:8080/ppsvm/) for

multiple participants to collaborate and complete the SVM-learning task in an

efficient and privacy-preserving manner.

PMCID: PMC3540462

PMID: 23304414 [Indexed for MEDLINE]

1492. AMIA Jt Summits Transl Sci Proc. 2012;2012:1-10. Epub 2012 Mar 19.

From sequencer to supercomputer: an automatic pipeline for managing and

processing next generation sequencing data.

Camerlengo T(1), Ozer HG, Onti-Srinivasan R, Yan P, Huang T, Parvin J, Huang K.

Author information:

(1)Department of Biomedical Informatics, The Ohio State University, Columbus,

Ohio, USA;

Next Generation Sequencing is highly resource intensive. NGS Tasks related to

data processing, management and analysis require high-end computing servers or

even clusters. Additionally, processing NGS experiments requires suitable storage

space and significant manual interaction. At The Ohio State University's

Biomedical Informatics Shared Resource, we designed and implemented a scalable

architecture to address the challenges associated with the resource intensive

nature of NGS secondary analysis built around Illumina Genome Analyzer II

sequencers and Illumina's Gerald data processing pipeline. The software

infrastructure includes a distributed computing platform consisting of a LIMS

called QUEST (http://bisr.osumc.edu), an Automation Server, a computer cluster

for processing NGS pipelines, and a network attached storage device expandable up

to 40TB. The system has been architected to scale to multiple sequencers without

requiring additional computing or labor resources. This platform provides

demonstrates how to manage and automate NGS experiments in an institutional or

core facility setting.

PMCID: PMC3392054

PMID: 22779037

1493. Bioinformatics. 2012 Jan 1;28(1):91-7. doi: 10.1093/bioinformatics/btr624. Epub

2011 Nov 15.

Protein subcellular localization of fluorescence imagery using spatial and

transform domain features.

Tahir M(1), Khan A, Majid A.

Author information:

(1)Department of Computer and Information Sciences, PIEAS, Islamabad, Pakistan.

MOTIVATION: Subcellular localization of proteins is one of the most significant

characteristics of living cells. Prediction of protein subcellular locations is

crucial to the understanding of various protein functions. Therefore, an

accurate, computationally efficient and reliable prediction system is required.

RESULTS: In this article, the predictions of various Support Vector Machine (SVM)

models have been combined through majority voting. The proposed ensemble

SVM-SubLoc has achieved the highest success rates of 99.7% using hybrid features

of Haralick textures and local binary patterns (HarLBP), 99.4% using hybrid

features of Haralick textures and Local Ternary Patterns (HarLTP). In addition,

SVM-SubLoc has yielded 99.0% accuracy using only local ternary patterns (LTPs)

based features. The dimensionality of HarLBP feature vector is 581 compared with

78 and 52 for HarLTP and LTPs, respectively. Hence, SVM-SubLoc in conjunction

with LTPs is fast, sufficiently accurate and simple predictive system. The

proposed SVM-SubLoc approach thus provides superior prediction performance using

the reduced feature space compared with existing approaches.

AVAILABILITY: A web server accompanying the proposed prediction scheme is

available at http://111.68.99.218/ SVM-SubLoc

CONTACT: asif@pieas.edu.pk; khan.asifullah@gmail.com

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr624

PMID: 22088847 [Indexed for MEDLINE]

1494. Bioinformatics. 2012 Jan 1;28(1):56-62. doi: 10.1093/bioinformatics/btr614. Epub

2011 Nov 8.

Epigenetic priors for identifying active transcription factor binding sites.

Cuellar-Partida G(1), Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL.

Author information:

(1)Institute for Molecular Bioscience, The University of Queensland, Brisbane QLD

4072, Australia.

MOTIVATION: Accurate knowledge of the genome-wide binding of transcription

factors in a particular cell type or under a particular condition is necessary

for understanding transcriptional regulation. Using epigenetic data such as

histone modification and DNase I, accessibility data has been shown to improve

motif-based in silico methods for predicting such binding, but this approach has

not yet been fully explored.

RESULTS: We describe a probabilistic method for combining one or more tracks of

epigenetic data with a standard DNA sequence motif model to improve our ability

to identify active transcription factor binding sites (TFBSs). We convert each

data type into a position-specific probabilistic prior and combine these priors

with a traditional probabilistic motif model to compute a log-posterior odds

score. Our experiments, using histone modifications H3K4me1, H3K4me3, H3K9ac and

H3K27ac, as well as DNase I sensitivity, show conclusively that the log-posterior

odds score consistently outperforms a simple binary filter based on the same

data. We also show that our approach performs competitively with a more complex

method, CENTIPEDE, and suggest that the relative simplicity of the log-posterior

odds scoring method makes it an appealing and very general method for identifying

functional TFBSs on the basis of DNA and epigenetic evidence.

AVAILABILITY AND IMPLEMENTATION: FIMO, part of the MEME Suite software toolkit,

now supports log-posterior odds scoring using position-specific priors for motif

search. A web server and source code are available at http://meme.nbcr.net.

Utilities for creating priors are at

http://research.imb.uq.edu.au/t.bailey/SD/Cuellar2011.

CONTACT: t.bailey@uq.edu.au

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr614

PMCID: PMC3244768

PMID: 22072382 [Indexed for MEDLINE]

1495. Bioinformatics. 2012 Jan 1;28(1):32-9. doi: 10.1093/bioinformatics/btr611. Epub

2011 Nov 7.

A novel structural position-specific scoring matrix for the prediction of protein

secondary structures.

Li D(1), Li T, Cong P, Xiong W, Sun J.

Author information:

(1)Department of Chemistry, Tongji University, Shanghai 200092, China.

MOTIVATION: The precise prediction of protein secondary structure is of key

importance for the prediction of 3D structure and biological function. Although

the development of many excellent methods over the last few decades has allowed

the achievement of prediction accuracies of up to 80%, progress seems to have

reached a bottleneck, and further improvements in accuracy have proven difficult.

RESULTS: We propose for the first time a structural position-specific scoring

matrix (SPSSM), and establish an unprecedented database of 9 million sequences

and their SPSSMs. This database, when combined with a purpose-designed BLAST

tool, provides a novel prediction tool: SPSSMPred. When the SPSSMPred was

validated on a large dataset (10,814 entries), the Q3 accuracy of the protein

secondary structure prediction was 93.4%. Our approach was tested on the two

latest EVA sets; accuracies of 82.7 and 82.0% were achieved, far higher than can

be achieved using other predictors. For further evaluation, we tested our

approach on newly determined sequences (141 entries), and obtained an accuracy of

89.6%. For a set of low-homology proteins (40 entries), the SPSSMPred still

achieved a Q3 value of 84.6%.

AVAILABILITY: The SPSSMPred server is available at

http://cal.tongji.edu.cn/SPSSMPred/

CONTACT: lith@tongji.edu.cn

DOI: 10.1093/bioinformatics/btr611

PMID: 22065541 [Indexed for MEDLINE]

1496. Bioinformatics. 2012 Jan 1;28(1):130-1. doi: 10.1093/bioinformatics/btr604. Epub

2011 Nov 3.

AMPA: an automated web server for prediction of protein antimicrobial regions.

Torrent M(1), Di Tommaso P, Pulido D, Nogués MV, Notredame C, Boix E, Andreu D.

Author information:

(1)Department of Experimental and Health Sciences, Universitat Pompeu Fabra,

Barcelona Biomedical Research Park, Dr Aiguader 88, Barcelona, Spain.

SUMMARY: AMPA is a web application for assessing the antimicrobial domains of

proteins, with a focus on the design on new antimicrobial drugs. The application

provides fast discovery of antimicrobial patterns in proteins that can be used to

develop new peptide-based drugs against pathogens. Results are shown in a

user-friendly graphical interface and can be downloaded as raw data for later

examination.

AVAILABILITY: AMPA is freely available on the web at

http://tcoffee.crg.cat/apps/ampa. The source code is also available in the web.

CONTACT: marc.torrent@upf.edu; david.andreu@upf.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr604

PMID: 22053077 [Indexed for MEDLINE]

1497. Bioinformation. 2012;8(25):1283-5. doi: 10.6026/97320630081283. Epub 2012 Dec 19.

StreptomycesInforSys: A web-enabled information repository.

Jain CK(1), Gupta V, Gupta A, Gupta S, Wadhwa G, Sharma SK, Sarethy IP.

Author information:

(1)Department of Biotechnology, Jaypee Institute of Information Technology, A-10,

Sector 62, Noida- 201307, India.

Members of Streptomyces produce 70% of natural bioactive products. There is

considerable amount of information available based on polyphasic approach for

classification of Streptomyces. However, this information based on phenotypic,

genotypic and bioactive component production profiles is crucial for

pharmacological screening programmes. This is scattered across various journals,

books and other resources, many of which are not freely accessible. The designed

database incorporates polyphasic typing information using combinations of search

options to aid in efficient screening of new isolates. This will help in the

preliminary categorization of appropriate groups. It is a free relational

database compatible with existing operating systems. A cross platform technology

with XAMPP Web server has been used to develop, manage, and facilitate the user

query effectively with database support. Employment of PHP, a

platform-independent scripting language, embedded in HTML and the database

management software MySQL will facilitate dynamic information storage and

retrieval. The user-friendly, open and flexible freeware (PHP, MySQL and Apache)

is foreseen to reduce running and maintenance cost.AVAILABILITY:

www.sis.biowaves.org.

DOI: 10.6026/97320630081283

PMCID: PMC3532016

PMID: 23275736

1498. Bioinformation. 2012;8(12):586-8. doi: 10.6026/97320630008586. Epub 2012 Jun 28.

Universal fingerprinting chip server.

Casique-Almazán J(1), Larios-Serrato V, Olguín-Ruíz GE, Sánchez-Vallejo CJ,

Maldonado-Rodríguez R, Méndez-Tenorio A.

Author information:

(1)Laboratory of Biotechnology and Genomic Bioinformatics, Department of

Biochemistry, National School of Biological Sciences, National Polytechnic

Institute, CP 11340, Mexico City, Mexico.

The Virtual Hybridization approach predicts the most probable hybridization sites

across a target nucleic acid of known sequence, including both perfect and

mismatched pairings. Potential hybridization sites, having a user-defined minimum

number of bases that are paired with the oligonucleotide probe, are first

identified. Then free energy values are evaluated for each potential

hybridization site, and if it has a calculated free energy of equal or higher

negative value than a user-defined free energy cut-off value, it is considered as

a site of high probability of hybridization. The Universal Fingerprinting Chip

Applications Server contains the software for visualizing predicted hybridization

patterns, which yields a simulated hybridization fingerprint that can be compared

with experimentally derived fingerprints or with a virtual fingerprint arising

from a different sample.AVAILABILITY: The database is available for free at

http://bioinformatica.homelinux.org/UFCVH/

DOI: 10.6026/97320630008586

PMCID: PMC3398781

PMID: 22829736

1499. Bioinformation. 2012;8(4):203-5. doi: 10.6026/97320630008203. Epub 2012 Feb 28.

GIST: Genomic island suite of tools for predicting genomic islands in genomic

sequences.

Hasan MS, Liu Q, Wang H, Fazekas J, Chen B, Che D.

Genomic Islands (GIs) are genomic regions that are originally from other

organisms, through a process known as Horizontal Gene Transfer (HGT). Detection

of GIs plays a significant role in biomedical research since such align genomic

regions usually contain important features, such as pathogenic genes. We have

developed a use friendly graphic user interface, Genomic Island Suite of Tools

(GIST), which is a platform for scientific users to predict GIs. This software

package includes five commonly used tools, AlienHunter, IslandPath, Colombo

SIGI-HMM, INDeGenIUS and Pai-Ida. It also includes an optimization program EGID

that ensembles the result of existing tools for more accurate prediction. The

tools in GIST can be used either separately or sequentially. GIST also includes a

downloadable feature that facilitates collecting the input genomes automatically

from the FTP server of the National Center for Biotechnology Information (NCBI).

GIST was implemented in Java, and was compiled and executed on Linux/Unix

operating systems.AVAILABILITY: The database is available for free at

http://www5.esu.edu/cpsc/bioinfo/software/GIST.

DOI: 10.6026/97320630008203

PMCID: PMC3302003

PMID: 22419842

1500. BMC Bioinformatics. 2012;13 Suppl 17:S7. doi: 10.1186/1471-2105-13-S17-S7. Epub

2012 Dec 13.

Sanjeevini: a freely accessible web-server for target directed lead molecule

discovery.

Jayaram B(1), Singh T, Mukherjee G, Mathur A, Shekhar S, Shekhar V.

Author information:

(1)Department of Chemistry, Indian Institute of Technology, Hauz Khas, New

Delhi-110016, India. bjayaram@chemistry.iitd.ac.in

BACKGROUND: Computational methods utilizing the structural and functional

information help to understand specific molecular recognition events between the

target biomolecule and candidate hits and make it possible to design improved

lead molecules for the target.

RESULTS: Sanjeevini represents a massive on-going scientific endeavor to provide

to the user, a freely accessible state of the art software suite for protein and

DNA targeted lead molecule discovery. It builds in several features, including

automated detection of active sites, scanning against a million compound library

for identifying hit molecules, all atom based docking and scoring and various

other utilities to design molecules with desired affinity and specificity against

biomolecular targets. Each of the modules is thoroughly validated on a large

dataset of protein/DNA drug targets.

CONCLUSIONS: The article presents Sanjeevini, a freely accessible user friendly

web-server, to aid in drug discovery. It is implemented on a tera flop cluster

and made accessible via a web-interface at

http://www.scfbio-iitd.res.in/sanjeevini/sanjeevini.jsp. A brief description of

various modules, their scientific basis, validation, and how to use the server to

develop in silico suggestions of lead molecules is provided.

DOI: 10.1186/1471-2105-13-S17-S7

PMCID: PMC3521208

PMID: 23282245 [Indexed for MEDLINE]

1501. BMC Bioinformatics. 2012;13 Suppl 17:S22. doi: 10.1186/1471-2105-13-S17-S22. Epub

2012 Dec 13.

Personalized cloud-based bioinformatics services for research and education: use

cases and the elasticHPC package.

El-Kalioby M(1), Abouelhoda M, Krüger J, Giegerich R, Sczyrba A, Wall DP,

Tonellato P.

Author information:

(1)Centre for Informatics Sciences, Nile University, Giza, Egypt.

mohamed.elkalioby@nileu.edu.eg

BACKGROUND: Bioinformatics services have been traditionally provided in the form

of a web-server that is hosted at institutional infrastructure and serves

multiple users. This model, however, is not flexible enough to cope with the

increasing number of users, increasing data size, and new requirements in terms

of speed and availability of service. The advent of cloud computing suggests a

new service model that provides an efficient solution to these problems, based on

the concepts of "resources-on-demand" and "pay-as-you-go". However, cloud

computing has not yet been introduced within bioinformatics servers due to the

lack of usage scenarios and software layers that address the requirements of the

bioinformatics domain.

RESULTS: In this paper, we provide different use case scenarios for providing

cloud computing based services, considering both the technical and financial

aspects of the cloud computing service model. These scenarios are for individual

users seeking computational power as well as bioinformatics service providers

aiming at provision of personalized bioinformatics services to their users. We

also present elasticHPC, a software package and a library that facilitates the

use of high performance cloud computing resources in general and the

implementation of the suggested bioinformatics scenarios in particular. Concrete

examples that demonstrate the suggested use case scenarios with whole

bioinformatics servers and major sequence analysis tools like BLAST are

presented. Experimental results with large datasets are also included to show the

advantages of the cloud model.

CONCLUSIONS: Our use case scenarios and the elasticHPC package are steps towards

the provision of cloud based bioinformatics services, which would help in

overcoming the data challenge of recent biological research. All resources

related to elasticHPC and its web-interface are available at

http://www.elasticHPC.org.

DOI: 10.1186/1471-2105-13-S17-S22

PMCID: PMC3521398

PMID: 23281941 [Indexed for MEDLINE]

1502. BMC Genomics. 2012;13 Suppl 7:S8. doi: 10.1186/1471-2164-13-S7-S8. Epub 2012 Dec

13.

Helminth secretome database (HSD): a collection of helminth excretory/secretory

proteins predicted from expressed sequence tags (ESTs).

Garg G(1), Ranganathan S.

Author information:

(1)Department of Chemistry and Biomolecular Sciences and ARC Centre of Excellence

in Bioinformatics, Macquarie University, Sydney NSW 2109, Australia.

BACKGROUND: Helminths are important socio-economic organisms, responsible for

causing major parasitic infections in humans, other animals and plants. These

infections impose a significant public health and economic burden globally.

Exceptionally, some helminth organisms like Caenorhabditis elegans are

free-living in nature and serve as model organisms for studying parasitic

infections. Excretory/secretory proteins play an important role in parasitic

helminth infections which make these proteins attractive targets for therapeutic

use. In the case of helminths, large volume of expressed sequence tags (ESTs) has

been generated to understand parasitism at molecular level and for predicting

excretory/secretory proteins for developing novel strategies to tackle parasitic

infections. However, mostly predicted ES proteins are not available for further

analysis and there is no repository available for such predicted ES proteins.

Furthermore, predictions have, in the main, focussed on classical secretory

pathways while it is well established that helminth parasites also utilise

non-classical secretory pathways.

RESULTS: We developed a free Helminth Secretome Database (HSD), which serves as a

repository for ES proteins predicted using classical and non-classical secretory

pathways, from EST data for 78 helminth species (64 nematodes, 7 trematodes and 7

cestodes) ranging from parasitic to free-living organisms. Approximately 0.9

million ESTs compiled from the largest EST database, dbEST were cleaned,

assembled and analysed by different computational tools in our bioinformatics

pipeline and predicted ES proteins were submitted to HSD.

CONCLUSION: We report the large-scale prediction and analysis of classically and

non-classically secreted ES proteins from diverse helminth organisms. All the

Unigenes (contigs and singletons) and excretory/secretory protein datasets

generated from this analysis are freely available. A BLAST server is available at

http://estexplorer.biolinfo.org/hsd, for checking the sequence similarity of new

protein sequences against predicted helminth ES proteins.

DOI: 10.1186/1471-2164-13-S7-S8

PMCID: PMC3546426

PMID: 23281827 [Indexed for MEDLINE]

1503. BMC Genomics. 2012;13 Suppl 7:S15. doi: 10.1186/1471-2164-13-S7-S15. Epub 2012

Dec 13.

A computational tool for the design of live attenuated virus vaccine based on

microRNA-mediated gene silencing.

Wichadakul D(1), Mhuantong W, Jongkaewwattana A, Ingsriswang S.

Author information:

(1)Information Systems Laboratory, National Center for Genetic Engineering and

Biotechnology (BIOTEC),113 Thailand Science Park, Phaholyothin Road, Klong 1,

Klong Luang, Pathumthani, 12120, Thailand. duangdao.wic@biotec.or.th

BACKGROUND: The microRNA-based gene-silencing machinery has been recognized as a

promising approach to control viral replication and used for improving safety for

the live attenuated virus vaccines. The effective host microRNA response elements

(MREs) have been incorporated into a virus sequence mainly based on the

experimental trials for identifying both microRNA binding sites and effective

mutations. The design of MREs for viral genomes or with multiple host microRNAs

of interest, then, will be time and cost consuming.

RESULTS: In this paper, we introduced a computational flow that could be used to

design MREs of human microRNAs within Influenza A H1N1 virus gene segments. The

main steps of the flow includes locating possible binding sites; MREs, of human

microRNAs within the viral sequences using a miRNA target prediction tool

(miranda), performing various mutations among mismatched binding positions,

calculating the binding energy, score, identity, and the effects of changed

physical properties of amino acids according to the changed bases in RNA level,

and prioritizing the mutated binding sites. The top ranked MREs of human microRNA

hsa-miR-93 is consistent with previous literature while other results waited to

be experimentally verified. To make the computational flow easily accessible by

virologists, we also developed MicroLive, a web server version of the MRE design

flow together with the database of miranda-predicted MREs within gene sequences

of seven RNA viruses including Influenza A, dengue, hepatitis C, measles, mumps,

poliovirus, and rabies. Users may design MREs of specific human microRNAs for

their input viral sequences using MRE design tool or optimize the

miranda-predicted MREs of seven viruses available on the system. Also, users

could design varied number of MREs for multiple human microRNAs to modulate the

degree of live vaccine attenuation and reduce the likelihood of escape mutants.

CONCLUSIONS: The computational design of MREs helps reduce time and cost for

experimental trials. While the flow was demonstrated using human microRNAs and

Influenza A H1N1 virus, it could be flexibly applied to other hosts (e.g.,

animals) and viruses of interest for constructing host-specific live attenuated

vaccines. Also, it could be deployed for engineering tissue-specific oncolytic

viruses in cancer virotherapeutics. The MicroLive web server is freely accessible

at http://www.biotec.or.th/isl/microlive.

DOI: 10.1186/1471-2164-13-S7-S15

PMCID: PMC3521223

PMID: 23281624 [Indexed for MEDLINE]

1504. BMC Genomics. 2012;13 Suppl 7:S6. doi: 10.1186/1471-2164-13-S7-S6. Epub 2012 Dec

13.

snpTree--a web-server to identify and construct SNP trees from whole genome

sequence data.

Leekitcharoenphon P(1), Kaas RS, Thomsen MC, Friis C, Rasmussen S, Aarestrup FM.

Author information:

(1)National Food Institute, Technical University of Denmark, 2800 Kgs Lyngby,

Denmark. pile@food.dtu.dk

BACKGROUND: The advances and decreasing economical cost of whole genome

sequencing (WGS), will soon make this technology available for routine infectious

disease epidemiology. In epidemiological studies, outbreak isolates have very

little diversity and require extensive genomic analysis to differentiate and

classify isolates. One of the successfully and broadly used methods is analysis

of single nucletide polymorphisms (SNPs). Currently, there are different tools

and methods to identify SNPs including various options and cut-off values.

Furthermore, all current methods require bioinformatic skills. Thus, we lack a

standard and simple automatic tool to determine SNPs and construct phylogenetic

tree from WGS data.

RESULTS: Here we introduce snpTree, a server for online-automatic SNPs analysis.

This tool is composed of different SNPs analysis suites, perl and python scripts.

snpTree can identify SNPs and construct phylogenetic trees from WGS as well as

from assembled genomes or contigs. WGS data in fastq format are aligned to

reference genomes by BWA while contigs in fasta format are processed by Nucmer.

SNPs are concatenated based on position on reference genome and a tree is

constructed from concatenated SNPs using FastTree and a perl script. The online

server was implemented by HTML, Java and python script.The server was evaluated

using four published bacterial WGS data sets (V. cholerae, S. aureus CC398, S.

Typhimurium and M. tuberculosis). The evaluation results for the first three

cases was consistent and concordant for both raw reads and assembled genomes. In

the latter case the original publication involved extensive filtering of SNPs,

which could not be repeated using snpTree.

CONCLUSIONS: The snpTree server is an easy to use option for rapid standardised

and automatic SNP analysis in epidemiological studies also for users with limited

bioinformatic experience. The web server is freely accessible at

http://www.cbs.dtu.dk/services/snpTree-1.0/.

DOI: 10.1186/1471-2164-13-S7-S6

PMCID: PMC3521233

PMID: 23281601 [Indexed for MEDLINE]

1505. Comput Biol Med. 2012 Jan;42(1):1-7. doi: 10.1016/j.compbiomed.2011.10.001. Epub

2011 Oct 29.

Predicting miRNA-mediated gene silencing mode based on miRNA-target duplex

features.

Song X(1), Cheng L, Zhou T, Guo X, Zhang X, Chen YP, Han P, Sha J.

Author information:

(1)Department of Biomedical Engineering, Nanjing University of Aeronautics and

Astronautics, Nanjing, China.

There are two main mechanisms of miRNA-mediated gene silencing: either mRNA

degradation or translational repression. However, the precise mechanism of target

mRNAs regulated by miRNA remains unclear. As a complementary approach to

experiment, a computational method was proposed to recognize the mechanism of

miRNA-mediated gene silencing in human. We have analyzed extensive features

correlated with miRNA-mediated silencing mechanism of mRNA. It is found that, the

duplex structure, the number of binding sites and the structural accessibility of

target site region are effective factors in determining whether a target mRNA is

cleaved or only translationally inhibited. An SVM-based classifier was

constructed to predict the regulation mode of miRNA based on these informative

features. The results indicated that the approach proposed is effective in

distinguishing whether a target mRNA is cleaved or translationally inhibited in

human. Furthermore, the web server microDoR (http://reprod.njmu.edu.cn/microdor)

has been developed and is freely available for users.

Crown Copyright © 2011. Published by Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiomed.2011.10.001

PMID: 22041293 [Indexed for MEDLINE]

1506. J Biomol Struct Dyn. 2012;29(6):659-70. doi: 10.1080/07391102.2011.672630.

Structural dynamics of full-length retroviral integrase: a molecular dynamics

analysis.

Balasubramanian S(1), Rajagopalan M, Ramaswamy A.

Author information:

(1)Centre for Bioinformatics, School of Life Sciences, Pondicherry University, R.

V. Nagar, Kalapet, Puducherry, India.

HIV integrase catalyzes the integration between host and viral DNA and is

considered as an interesting target for treating HIV. Knowledge of the complete

structure of integrase is inevitable to describe the communicative inter-domain

interactions affecting the HIV integration and disintegration process and hence

the study on full-length integrase turns out to be an essential task. In this

investigation, a structure of full-length integrase is designed to analyze the

global dynamics of integrase dimer and monomers (with and without the C-terminal,

270-288 amino acids) for a period of 20 ns. The molecular dynamics analysis and

the subsequent DynDom analysis reveal (i) a stable dynamics of dimeric CCD and

NTD domains and (ii) CCD-α11-mediated rotational-cum-translational CTD motion as

the functional dynamics of IN dimer. This observation supports that (i)

aggregation enhances the integrase activity and (ii) flexible CTD for its cis and

trans coordination with CCD. The role of C-loop over the dynamics of integrase is

also explored, which unveils that the spatial arrangement of integrase domains is

changed during dynamics in the absence of C-loop. In essence, here we report a

C-loop-dependent structural dynamics of integrase and the active dynamics of

integrase in dimer. Further studies on C-loop sensing mechanism and the

multimerization of integrase would provide insight into HIV integration and

disintegration processes. Supplementary material. Movies generated from molecular

dynamics trajectory showing the CTD dynamics of IN structures (monomers with &

without C-loop and dimer) are linked online to this article. The remaining

supplementary data can be downloaded from the author's server at the URL

http://ramutha.bicpu.edu.in .

DOI: 10.1080/07391102.2011.672630

PMID: 22545997 [Indexed for MEDLINE]

1507. Methods Mol Biol. 2012;882:605-33. doi: 10.1007/978-1-61779-842-9\_33.

IMGT/DomainGapAlign: the IMGT® tool for the analysis of IG, TR, MH, IgSF, and

MhSF domain amino acid polymorphism.

Ehrenmann F(1), Lefranc MP.

Author information:

(1)IMGT®, The International ImMunoGeneTics Information System®, Université

Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de

Génétique Humaine (IGH), UPR CNRS 1142, Montpellier, France.

IMGT/DomainGapAlign is the online tool of IMGT(®), the international

ImMunoGeneTics information system(®), for the analysis of amino acid sequences

and two-dimensional (2D) structures of domains. IMGT/DomainGapAlign allows the

analysis of the closest variable (V) and constant (C) domains of immunoglobulins

(IG) or antibodies, T cell receptors (TR), and immunoglobulin superfamily (IgSF)

proteins, and of the groove (G) domains of major histocompatibility (MH; in

humans, HLA for human leukocyte antigen) and MH superfamily proteins.

IMGT/DomainGapAlign aligns the user own sequences against the IMGT domain

reference directory, displays amino acid changes, creates IMGT gaps, and delimits

the domain strands and loops (and helix for G domain) according to the IMGT

unique numbering. IMGT/DomainGapAlign is coupled to the IMGT/Collier-de-Perles

tool that draws standardized IMGT Colliers de Perles. The analysis is based on

the IMGT-ONTOLOGY concepts of identification, classification, description, and

numerotation generated from the axioms of the Formal IMGT-ONTOLOGY or

IMGT-Kaleidoscope. IMGT/DomainGapAlign provides an invaluable help for antibody

engineering and antibody humanization as it precisely defines the standardized

framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT) to

be grafted. IMGT/DomainGapAlign is freely available at http://www.imgt.org.

DOI: 10.1007/978-1-61779-842-9\_33

PMID: 22665257 [Indexed for MEDLINE]

1508. Methods Mol Biol. 2012;819:13-27. doi: 10.1007/978-1-61779-465-0\_2.

Analysis of protein binding sites by computational solvent mapping.

Hall DR(1), Kozakov D, Vajda S.

Author information:

(1)Department of Biomedical Engineering, Biomolecular Engineering Research

Center, Boston University, Boston, MA, USA. drhall@bu.edu

Computational solvent mapping globally samples the surface of target proteins

using molecular probes-small molecules or functional groups-to identify

potentially favorable binding positions. The method is based on X-ray and NMR

screening studies showing that the binding sites of proteins also bind a large

variety of fragment-sized molecules. We have developed the multistage mapping

algorithm FTMap (available as a server at http://ftmap.bu.edu/ ) based on the

fast Fourier transform (FFT) correlation approach. Identifying regions of low

free energy rather than individual low energy conformations, FTMap reproduces the

available experimental mapping results. Applications to a variety of proteins

show that the probes always cluster in important subsites of the binding site,

and the amino acid residues that interact with many probes also bind the specific

ligands of the protein. The "consensus" sites at which a number of different

probes cluster are likely to be "druggable" sites, capable of binding drug-size

ligands with high affinity. Due to its sensitivity to conformational changes, the

method can also be used for comparing the binding sites in different structures

of a protein.

DOI: 10.1007/978-1-61779-465-0\_2

PMCID: PMC3526383

PMID: 22183527 [Indexed for MEDLINE]

1509. Methods Mol Biol. 2012;804:107-30. doi: 10.1007/978-1-61779-361-5\_7.

Predicting metabolic pathways by sub-network extraction.

Faust K(1), van Helden J.

Author information:

(1)Department of Applied Biological Sciences (DBIT), Vrije Universiteit Brussel,

Bruxelles, Belgium. kfaust@ulb.ac.be

Various methods result in groups of functionally related genes obtained from

genomes (operons, regulons, syntheny groups, and phylogenetic profiles),

transcriptomes (co-expression groups) and proteomes (modules of interacting

proteins). When such groups contain two or more enzyme-coding genes, graph

analysis methods can be applied to extract a metabolic pathway that interconnects

them. We describe here the way to use the Pathway extraction tool available on

the NeAT Web server ( http://rsat.ulb.ac.be/neat/ ) to piece together the

metabolic pathway from a group of associated, enzyme-coding genes. The tool

identifies the reactions that can be catalyzed by the products of the query genes

(seed reactions), and applies sub-graph extraction algorithms to extract from a

metabolic network a sub-network that connects the seed reactions. This

sub-network represents the predicted metabolic pathway. We describe here the

pathway prediction process in a step-by-step way, give hints about the main

parametric choices, and illustrate how this tool can be used to extract metabolic

pathways from bacterial genomes, on the basis of two study cases: the

isoleucine-valine operon in Escherichia coli and a predicted operon in

Cupriavidus (Ralstonia) metallidurans.

DOI: 10.1007/978-1-61779-361-5\_7

PMID: 22144151 [Indexed for MEDLINE]

1510. Nucleic Acids Res. 2012 Jan;40(Database issue):D210-5. doi: 10.1093/nar/gkr1175.

Epub 2011 Dec 1.

NONCODE v3.0: integrative annotation of long noncoding RNAs.

Bu D(1), Yu K, Sun S, Xie C, Skogerbø G, Miao R, Xiao H, Liao Q, Luo H, Zhao G,

Zhao H, Liu Z, Liu C, Chen R, Zhao Y.

Author information:

(1)Bioinformatics Research Group, Key Laboratory of Intelligent Information

Processing, Advanced Computer Research Center, Institute of Computing Technology,

Chinese Academy of Sciences, Beijing, PR China.

Facilitated by the rapid progress of high-throughput sequencing technology, a

large number of long noncoding RNAs (lncRNAs) have been identified in mammalian

transcriptomes over the past few years. LncRNAs have been shown to play key roles

in various biological processes such as imprinting control, circuitry controlling

pluripotency and differentiation, immune responses and chromosome dynamics.

Notably, a growing number of lncRNAs have been implicated in disease etiology.

With the increasing number of published lncRNA studies, the experimental data on

lncRNAs (e.g. expression profiles, molecular features and biological functions)

have accumulated rapidly. In order to enable a systematic compilation and

integration of this information, we have updated the NONCODE database

(http://www.noncode.org) to version 3.0 to include the first integrated

collection of expression and functional lncRNA data obtained from re-annotated

microarray studies in a single database. NONCODE has a user-friendly interface

with a variety of search or browse options, a local Genome Browser for

visualization and a BLAST server for sequence-alignment search. In addition,

NONCODE provides a platform for the ongoing collation of ncRNAs reported in the

literature. All data in NONCODE are open to users, and can be downloaded through

the website or obtained through the SOAP API and DAS services.

DOI: 10.1093/nar/gkr1175

PMCID: PMC3245065

PMID: 22135294 [Indexed for MEDLINE]

1511. Nucleic Acids Res. 2012 Jan;40(Database issue):D1137-43. doi: 10.1093/nar/gkr973.

Epub 2011 Nov 18.

The Human OligoGenome Resource: a database of oligonucleotide capture probes for

resequencing target regions across the human genome.

Newburger DE(1), Natsoulis G, Grimes S, Bell JM, Davis RW, Batzoglou S, Ji HP.

Author information:

(1)Biomedical Informatics Training Program, Division of Oncology, Department of

Medicine, Stanford University School of Medicine, Stanford, CA 94305, Stanford

Genome Technology Center, Stanford University, Palo Alto and Department of

Computer Science, Stanford University, Stanford, CA 94304, USA.

Recent exponential growth in the throughput of next-generation DNA sequencing

platforms has dramatically spurred the use of accessible and scalable targeted

resequencing approaches. This includes candidate region diagnostic resequencing

and novel variant validation from whole genome or exome sequencing analysis. We

have previously demonstrated that selective genomic circularization is a robust

in-solution approach for capturing and resequencing thousands of target human

genome loci such as exons and regulatory sequences. To facilitate the design and

production of customized capture assays for any given region in the human genome,

we developed the Human OligoGenome Resource (http://oligogenome.stanford.edu/).

This online database contains over 21 million capture oligonucleotide sequences.

It enables one to create customized and highly multiplexed resequencing assays of

target regions across the human genome and is not restricted to coding regions.

In total, this resource provides 92.1% in silico coverage of the human genome.

The online server allows researchers to download a complete repository of

oligonucleotide probes and design customized capture assays to target multiple

regions throughout the human genome. The website has query tools for selecting

and evaluating capture oligonucleotides from specified genomic regions.

DOI: 10.1093/nar/gkr973

PMCID: PMC3245143

PMID: 22102592 [Indexed for MEDLINE]

1512. Nucleic Acids Res. 2012 Jan;40(Database issue):D834-40. doi: 10.1093/nar/gkr997.

Epub 2011 Nov 18.

IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates

multiple types of conserved interactions for proteins.

Shoemaker BA(1), Zhang D, Tyagi M, Thangudu RR, Fong JH, Marchler-Bauer A, Bryant

SH, Madej T, Panchenko AR.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, 8600 Rockville Pike, Building 38A, Bethesda, MD

20894, USA.

We have recently developed the Inferred Biomolecular Interaction Server (IBIS)

and database, which reports, predicts and integrates different types of

interaction partners and locations of binding sites in proteins based on the

analysis of homologous structural complexes. Here, we highlight several new IBIS

features and options. The server's webpage is now redesigned to allow users

easier access to data for different interaction types. An entry page is added to

give a quick summary of available results and to now accept protein sequence

accessions. To elucidate the formation of protein complexes, not just binary

interactions, IBIS currently presents an expandable interaction network.

Previously, IBIS provided annotations for four different types of binding

partners: proteins, small molecules, nucleic acids and peptides; in the current

version a new protein-ion interaction type has been added. Several options

provide easy downloads of IBIS data for all Protein Data Bank (PDB) protein

chains and the results for each query. In this study, we show that about

one-third of all RefSeq sequences can be annotated with IBIS interaction partners

and binding sites. The IBIS server is available at

http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi and updated biweekly.

DOI: 10.1093/nar/gkr997

PMCID: PMC3245142

PMID: 22102591 [Indexed for MEDLINE]

1513. Nucleic Acids Res. 2012 Jan;40(Database issue):D615-20. doi: 10.1093/nar/gkr942.

Epub 2011 Nov 18.

DBETH: a Database of Bacterial Exotoxins for Human.

Chakraborty A(1), Ghosh S, Chowdhary G, Maulik U, Chakrabarti S.

Author information:

(1)Department of Structural Biology and Bioinformatics Division, Indian Institute

of Chemical Biology, Council for Scientific and Industrial Research, Jadavpur

University, Kolkata, WB 700 032, India.

Pathogenic bacteria produce protein toxins to survive in the hostile environments

defined by the host's defense systems and immune response. Recent progresses in

high-throughput genome sequencing and structure determination techniques have

contributed to a better understanding of mechanisms of action of the bacterial

toxins at the cellular and molecular levels leading to pathogenicity. It is fair

to assume that with time more and more unknown toxins will emerge not only by the

discovery of newer species but also due to the genetic rearrangement of existing

bacterial genomes. Hence, it is crucial to organize a systematic compilation and

subsequent analyses of the inherent features of known bacterial toxins. We

developed a Database for Bacterial ExoToxins (DBETH,

http://www.hpppi.iicb.res.in/btox/), which contains sequence, structure,

interaction network and analytical results for 229 toxins categorized within 24

mechanistic and activity types from 26 bacterial genuses. The main objective of

this database is to provide a comprehensive knowledgebase for human pathogenic

bacterial toxins where various important sequence, structure and physico-chemical

property based analyses are provided. Further, we have developed a prediction

server attached to this database which aims to identify bacterial toxin like

sequences either by establishing homology with known toxin sequences/domains or

by classifying bacterial toxin specific features using a support vector based

machine learning techniques.

DOI: 10.1093/nar/gkr942

PMCID: PMC3244994

PMID: 22102573 [Indexed for MEDLINE]

1514. Nucleic Acids Res. 2012 Jan;40(2):487-98. doi: 10.1093/nar/gkr629. Epub 2011 Sep

14.

Predicting coaxial helical stacking in RNA junctions.

Laing C(1), Wen D, Wang JT, Schlick T.

Author information:

(1)Department of Chemistry, Courant Institute of Mathematical Sciences, New York

University, 251 Mercer Street, New York, NY 10012, USA.

RNA junctions are important structural elements that form when three or more

helices come together in space in the tertiary structures of RNA molecules.

Determining their structural configuration is important for predicting RNA 3D

structure. We introduce a computational method to predict, at the secondary

structure level, the coaxial helical stacking arrangement in junctions, as well

as classify the junction topology. Our approach uses a data mining approach known

as random forests, which relies on a set of decision trees trained using length,

sequence and other variables specified for any given junction. The resulting

protocol predicts coaxial stacking within three- and four-way junctions with an

accuracy of 81% and 77%, respectively; the accuracy increases to 83% and 87%,

respectively, when knowledge from the junction family type is included. Coaxial

stacking predictions for the five to ten-way junctions are less accurate (60%)

due to sparse data available for training. Additionally, our application predicts

the junction family with an accuracy of 85% for three-way junctions and 74% for

four-way junctions. Comparisons with other methods, as well applications to

unsolved RNAs, are also presented. The web server Junction-Explorer to predict

junction topologies is freely available at:

http://bioinformatics.njit.edu/junction.

DOI: 10.1093/nar/gkr629

PMCID: PMC3258123

PMID: 21917853 [Indexed for MEDLINE]

1515. Nucleic Acids Res. 2012 Jan;40(Database issue):D370-6. doi: 10.1093/nar/gkr703.

Epub 2011 Sep 2.

OPM database and PPM web server: resources for positioning of proteins in

membranes.

Lomize MA(1), Pogozheva ID, Joo H, Mosberg HI, Lomize AL.

Author information:

(1)Department of Medicinal Chemistry, College of Pharmacy, University of

Michigan, 428 Church Street, Ann Arbor, MI 48109-1065 USA. almz@umich.edu

The Orientations of Proteins in Membranes (OPM) database is a curated web

resource that provides spatial positions of membrane-bound peptides and proteins

of known three-dimensional structure in the lipid bilayer, together with their

structural classification, topology and intracellular localization. OPM currently

contains more than 1200 transmembrane and peripheral proteins and peptides from

approximately 350 organisms that represent approximately 3800 Protein Data Bank

entries. Proteins are classified into classes, superfamilies and families and

assigned to 21 distinct membrane types. Spatial positions of proteins with

respect to the lipid bilayer are optimized by the PPM 2.0 method that accounts

for the hydrophobic, hydrogen bonding and electrostatic interactions of the

proteins with the anisotropic water-lipid environment described by the dielectric

constant and hydrogen-bonding profiles. The OPM database is freely accessible at

http://opm.phar.umich.edu. Data can be sorted, searched or retrieved using the

hierarchical classification, source organism, localization in different types of

membranes. The database offers downloadable coordinates of proteins and peptides

with membrane boundaries. A gallery of protein images and several visualization

tools are provided. The database is supplemented by the PPM server

(http://opm.phar.umich.edu/server.php) which can be used for calculating spatial

positions in membranes of newly determined proteins structures or theoretical

models.

DOI: 10.1093/nar/gkr703

PMCID: PMC3245162

PMID: 21890895 [Indexed for MEDLINE]

1516. Pac Symp Biocomput. 2012:164-75.

Functional annotation of intrinsically disordered domains by their amino acid

content using IDD Navigator.

Patil A(1), Teraguchi S, Dinh H, Nakai K, Standley DM.

Author information:

(1)Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo,

Japan. ashwini@hgc.jp

Function prediction of intrinsically disordered domains (IDDs) using sequence

similarity methods is limited by their high mutability and prevalence of low

complexity regions. We describe a novel method for identifying similar IDDs by a

similarity metric based on amino acid composition and identify significantly

overrepresented Gene Ontology (GO) and Pfam domain annotations within highly

similar IDDs. Applications and extensions of the proposed method are discussed,

in particular with respect to protein functional annotation. We test the

predicted annotations in a large-scale survey of IDDs in mouse and find that the

proposed method provides significantly greater protein coverage in terms of

function prediction than traditional sequence alignment methods like BLAST. As a

proof of concept we examined several disorder-containing proteins: GRA15 and

ROP16, both encoded in the parasitic protozoa T. gondii; Cyclon, a mostly

uncharacterized protein involved in the regulation of immune cell death; STIM1, a

protein essential for regulating calcium levels in the endoplasmic reticulum. We

show that the overrepresented GO terms are consistent with recently-reported

biological functions. We implemented the method in the web server IDD Navigator.

IDD Navigator is available at

http://sysimm.ifrec.osaka-u.ac.jp/disorder/beta.php.

PMID: 22174272 [Indexed for MEDLINE]

1517. PLoS Comput Biol. 2012;8(6):e1002567. doi: 10.1371/journal.pcbi.1002567. Epub

2012 Jun 28.

ProteinHistorian: tools for the comparative analysis of eukaryote protein origin.

Capra JA(1), Williams AG, Pollard KS.

Author information:

(1)J. David Gladstone Institutes, University of California, San Francisco,

California, United States of America. tony.capra@gladstone.ucsf.edu

The evolutionary history of a protein reflects the functional history of its

ancestors. Recent phylogenetic studies identified distinct evolutionary

signatures that characterize proteins involved in cancer, Mendelian disease, and

different ontogenic stages. Despite the potential to yield insight into the

cellular functions and interactions of proteins, such comparative phylogenetic

analyses are rarely performed, because they require custom algorithms. We

developed ProteinHistorian to make tools for performing analyses of protein

origins widely available. Given a list of proteins of interest, ProteinHistorian

estimates the phylogenetic age of each protein, quantifies enrichment for

proteins of specific ages, and compares variation in protein age with other

protein attributes. ProteinHistorian allows flexibility in the definition of

protein age by including several algorithms for estimating ages from different

databases of evolutionary relationships. We illustrate the use of

ProteinHistorian with three example analyses. First, we demonstrate that proteins

with high expression in human, compared to chimpanzee and rhesus macaque, are

significantly younger than those with human-specific low expression. Next, we

show that human proteins with annotated regulatory functions are significantly

younger than proteins with catalytic functions. Finally, we compare protein

length and age in many eukaryotic species and, as expected from previous studies,

find a positive, though often weak, correlation between protein age and length.

ProteinHistorian is available through a web server with an intuitive interface

and as a set of command line tools; this allows biologists and bioinformaticians

alike to integrate these approaches into their analysis pipelines.

ProteinHistorian's modular, extensible design facilitates the integration of new

datasets and algorithms. The ProteinHistorian web server, source code, and

pre-computed ages for 32 eukaryotic genomes are freely available under the GNU

public license at http://lighthouse.ucsf.edu/ProteinHistorian/.

DOI: 10.1371/journal.pcbi.1002567

PMCID: PMC3386163

PMID: 22761559 [Indexed for MEDLINE]

1518. PLoS One. 2012;7(12):e52901. doi: 10.1371/journal.pone.0052901. Epub 2012 Dec 28.

Assignment of EC numbers to enzymatic reactions with reaction difference

fingerprints.

Hu QN(1), Zhu H, Li X, Zhang M, Deng Z, Yang X, Deng Z.

Author information:

(1)Key Laboratory of Combinatorial Biosynthesis and Drug Discovery, Wuhan

University, Ministry of Education, Wuhan, P. R. China. qnhu@whu.edu.cn

The EC numbers represent enzymes and enzyme genes (genomic information), but they

are also utilized as identifiers of enzymatic reactions (chemical information).

In the present work (ECAssigner), our newly proposed reaction difference

fingerprints (RDF) are applied to assign EC numbers to enzymatic reactions. The

fingerprints of reactant molecules minus the fingerprints of product molecules

will generate reaction difference fingerprints, which are then used to calculate

reaction Euclidean distance, a reaction similarity measurement, of two reactions.

The EC number of the most similar training reaction will be assigned to an input

reaction. For 5120 balanced enzymatic reactions, the RDF with a fingerprint

length at 3 obtained at the sub-subclass, subclass, and main class level with

cross-validation accuracies of 83.1%, 86.7%, and 92.6% respectively. Compared

with three published methods, ECAssigner is the first fully automatic server for

EC number assignment. The EC assignment system (ECAssigner) is freely available

via: http://cadd.whu.edu.cn/ecassigner/.

DOI: 10.1371/journal.pone.0052901

PMCID: PMC3532301

PMID: 23285222 [Indexed for MEDLINE]

1519. PLoS One. 2012;7(12):e50506. doi: 10.1371/journal.pone.0050506. Epub 2012 Dec 19.

Using the fast fourier transform to accelerate the computational search for RNA

conformational switches.

Senter E(1), Sheikh S, Dotu I, Ponty Y, Clote P.

Author information:

(1)Biology Department, Boston College, Chestnut Hill, Massachusetts, United

States of America.

Using complex roots of unity and the Fast Fourier Transform, we design a new

thermodynamics-based algorithm, FFTbor, that computes the Boltzmann probability

that secondary structures differ by [Formula: see text] base pairs from an

arbitrary initial structure of a given RNA sequence. The algorithm, which runs in

quartic time O(n(4)) and quadratic space O(n(2)), is used to determine the

correlation between kinetic folding speed and the ruggedness of the energy

landscape, and to predict the location of riboswitch expression platform

candidates. A web server is available at

http://bioinformatics.bc.edu/clotelab/FFTbor/.

DOI: 10.1371/journal.pone.0050506

PMCID: PMC3526635

PMID: 23284639 [Indexed for MEDLINE]

1520. PLoS One. 2012;7(12):e51252. doi: 10.1371/journal.pone.0051252. Epub 2012 Dec 7.

CINPER: an interactive web system for pathway prediction for prokaryotes.

Mao X(1), Chen X, Zhang Y, Pangle S, Xu Y.

Author information:

(1)Computational Systems Biology Lab, Department of Biochemistry and Molecular

Biology and Institute of Bioinformatics, University of Georgia, Athens, Georgia,

United States of America.

We present a web-based network-construction system, CINPER (CSBL INteractive

Pathway BuildER), to assist a user to build a user-specified gene network for a

prokaryotic organism in an intuitive manner. CINPER builds a network model based

on different types of information provided by the user and stored in the system.

CINPER's prediction process has four steps: (i) collection of template networks

based on (partially) known pathways of related organism(s) from the SEED or

BioCyc database and the published literature; (ii) construction of an initial

network model based on the template networks using the P-Map program; (iii)

expansion of the initial model, based on the association information derived from

operons, protein-protein interactions, co-expression modules and phylogenetic

profiles; and (iv) computational validation of the predicted models based on gene

expression data. To facilitate easy applications, CINPER provides an interactive

visualization environment for a user to enter, search and edit relevant data and

for the system to display (partial) results and prompt for additional data.

Evaluation of CINPER on 17 well-studied pathways in the MetaCyc database shows

that the program achieves an average recall rate of 76% and an average precision

rate of 90% on the initial models; and a higher average recall rate at 87% and an

average precision rate at 28% on the final models. The reduced precision rate in

the final models versus the initial models reflects the reality that the final

models have large numbers of novel genes that have no experimental evidences and

hence are not yet collected in the MetaCyc database. To demonstrate the

usefulness of this server, we have predicted an iron homeostasis gene network of

Synechocystis sp. PCC6803 using the server. The predicted models along with the

server can be accessed at http://csbl.bmb.uga.edu/cinper/.

DOI: 10.1371/journal.pone.0051252

PMCID: PMC3517448

PMID: 23236458 [Indexed for MEDLINE]

1521. PLoS One. 2012;7(11):e50300. doi: 10.1371/journal.pone.0050300. Epub 2012 Nov 29.

PROSPER: an integrated feature-based tool for predicting protease substrate

cleavage sites.

Song J(1), Tan H, Perry AJ, Akutsu T, Webb GI, Whisstock JC, Pike RN.

Author information:

(1)Department of Biochemistry and Molecular Biology, Monash University,

Melbourne, Australia. Jiangning.Song@monash.edu

The ability to catalytically cleave protein substrates after synthesis is

fundamental for all forms of life. Accordingly, site-specific proteolysis is one

of the most important post-translational modifications. The key to understanding

the physiological role of a protease is to identify its natural substrate(s).

Knowledge of the substrate specificity of a protease can dramatically improve our

ability to predict its target protein substrates, but this information must be

utilized in an effective manner in order to efficiently identify protein

substrates by in silico approaches. To address this problem, we present PROSPER,

an integrated feature-based server for in silico identification of protease

substrates and their cleavage sites for twenty-four different proteases. PROSPER

utilizes established specificity information for these proteases (derived from

the MEROPS database) with a machine learning approach to predict protease

cleavage sites by using different, but complementary sequence and structure

characteristics. Features used by PROSPER include local amino acid sequence

profile, predicted secondary structure, solvent accessibility and predicted

native disorder. Thus, for proteases with known amino acid specificity, PROSPER

provides a convenient, pre-prepared tool for use in identifying protein

substrates for the enzymes. Systematic prediction analysis for the twenty-four

proteases thus far included in the database revealed that the features we have

included in the tool strongly improve performance in terms of cleavage site

prediction, as evidenced by their contribution to performance improvement in

terms of identifying known cleavage sites in substrates for these enzymes. In

comparison with two state-of-the-art prediction tools, PoPS and SitePrediction,

PROSPER achieves greater accuracy and coverage. To our knowledge, PROSPER is the

first comprehensive server capable of predicting cleavage sites of multiple

proteases within a single substrate sequence using machine learning techniques.

It is freely available at http://lightning.med.monash.edu.au/PROSPER/.

DOI: 10.1371/journal.pone.0050300

PMCID: PMC3510211

PMID: 23209700 [Indexed for MEDLINE]

1522. PLoS One. 2012;7(11):e49029. doi: 10.1371/journal.pone.0049029. Epub 2012 Nov 28.

Wiki-pi: a web-server of annotated human protein-protein interactions to aid in

discovery of protein function.

Orii N(1), Ganapathiraju MK.

Author information:

(1)Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh,

PA, USA.

Protein-protein interactions (PPIs) are the basis of biological functions.

Knowledge of the interactions of a protein can help understand its molecular

function and its association with different biological processes and pathways.

Several publicly available databases provide comprehensive information about

individual proteins, such as their sequence, structure, and function. There also

exist databases that are built exclusively to provide PPIs by curating them from

published literature. The information provided in these web resources is

protein-centric, and not PPI-centric. The PPIs are typically provided as lists of

interactions of a given gene with links to interacting partners; they do not

present a comprehensive view of the nature of both the proteins involved in the

interactions. A web database that allows search and retrieval based on biomedical

characteristics of PPIs is lacking, and is needed. We present Wiki-Pi (read

Wiki-π), a web-based interface to a database of human PPIs, which allows users to

retrieve interactions by their biomedical attributes such as their association to

diseases, pathways, drugs and biological functions. Each retrieved PPI is shown

with annotations of both of the participant proteins side-by-side, creating a

basis to hypothesize the biological function facilitated by the interaction.

Conceptually, it is a search engine for PPIs analogous to PubMed for scientific

literature. Its usefulness in generating novel scientific hypotheses is

demonstrated through the study of IGSF21, a little-known gene that was recently

identified to be associated with diabetic retinopathy. Using Wiki-Pi, we infer

that its association to diabetic retinopathy may be mediated through its

interactions with the genes HSPB1, KRAS, TMSB4X and DGKD, and that it may be

involved in cellular response to external stimuli, cytoskeletal organization and

regulation of molecular activity. The website also provides a wiki-like

capability allowing users to describe or discuss an interaction. Wiki-Pi is

available publicly and freely at http://severus.dbmi.pitt.edu/wiki-pi/.

DOI: 10.1371/journal.pone.0049029

PMCID: PMC3509123

PMID: 23209562 [Indexed for MEDLINE]

1523. PLoS One. 2012;7(11):e49040. doi: 10.1371/journal.pone.0049040. Epub 2012 Nov 26.

Predicting secretory proteins of malaria parasite by incorporating sequence

evolution information into pseudo amino acid composition via grey system model.

Lin WZ(1), Fang JA, Xiao X, Chou KC.

Author information:

(1)Information Science and technology School, Donghua University, Shanghai,

China.

The malaria disease has become a cause of poverty and a major hindrance to

economic development. The culprit of the disease is the parasite, which secretes

an array of proteins within the host erythrocyte to facilitate its own survival.

Accordingly, the secretory proteins of malaria parasite have become a logical

target for drug design against malaria. Unfortunately, with the increasing

resistance to the drugs thus developed, the situation has become more

complicated. To cope with the drug resistance problem, one strategy is to timely

identify the secreted proteins by malaria parasite, which can serve as potential

drug targets. However, it is both expensive and time-consuming to identify the

secretory proteins of malaria parasite by experiments alone. To expedite the

process for developing effective drugs against malaria, a computational predictor

called "iSMP-Grey" was developed that can be used to identify the secretory

proteins of malaria parasite based on the protein sequence information alone.

During the prediction process a protein sample was formulated with a 60D

(dimensional) feature vector formed by incorporating the sequence evolution

information into the general form of PseAAC (pseudo amino acid composition) via a

grey system model, which is particularly useful for solving complicated problems

that are lack of sufficient information or need to process uncertain information.

It was observed by the jackknife test that iSMP-Grey achieved an overall success

rate of 94.8%, remarkably higher than those by the existing predictors in this

area. As a user-friendly web-server, iSMP-Grey is freely accessible to the public

at http://www.jci-bioinfo.cn/iSMP-Grey. Moreover, for the convenience of most

experimental scientists, a step-by-step guide is provided on how to use the

web-server to get the desired results without the need to follow the complicated

mathematical equations involved in this paper.

DOI: 10.1371/journal.pone.0049040

PMCID: PMC3506597

PMID: 23189138 [Indexed for MEDLINE]

1524. PLoS One. 2012;7(11):e49334. doi: 10.1371/journal.pone.0049334. Epub 2012 Nov 8.

CREST--classification resources for environmental sequence tags.

Lanzén A(1), Jørgensen SL, Huson DH, Gorfer M, Grindhaug SH, Jonassen I, Øvreås

L, Urich T.

Author information:

(1)Department of Biology and Centre for Geobiology, University of Bergen, Bergen,

Norway. anders.lanzen@uni.no

Sequencing of taxonomic or phylogenetic markers is becoming a fast and efficient

method for studying environmental microbial communities. This has resulted in a

steadily growing collection of marker sequences, most notably of the

small-subunit (SSU) ribosomal RNA gene, and an increased understanding of

microbial phylogeny, diversity and community composition patterns. However, to

utilize these large datasets together with new sequencing technologies, a

reliable and flexible system for taxonomic classification is critical. We

developed CREST (Classification Resources for Environmental Sequence Tags), a set

of resources and tools for generating and utilizing custom taxonomies and

reference datasets for classification of environmental sequences. CREST uses an

alignment-based classification method with the lowest common ancestor algorithm.

It also uses explicit rank similarity criteria to reduce false positives and

identify novel taxa. We implemented this method in a web server, a command line

tool and the graphical user interfaced program MEGAN. Further, we provide the SSU

rRNA reference database and taxonomy SilvaMod, derived from the publicly

available SILVA SSURef, for classification of sequences from bacteria, archaea

and eukaryotes. Using cross-validation and environmental datasets, we compared

the performance of CREST and SilvaMod to the RDP Classifier. We also utilized

Greengenes as a reference database, both with CREST and the RDP Classifier. These

analyses indicate that CREST performs better than alignment-free methods with

higher recall rate (sensitivity) as well as precision, and with the ability to

accurately identify most sequences from novel taxa. Classification using SilvaMod

performed better than with Greengenes, particularly when applied to environmental

sequences. CREST is freely available under a GNU General Public License (v3) from

http://apps.cbu.uib.no/crest and http://lcaclassifier.googlecode.com.

DOI: 10.1371/journal.pone.0049334

PMCID: PMC3493522

PMID: 23145153 [Indexed for MEDLINE]

1525. PLoS One. 2012;7(10):e47843. doi: 10.1371/journal.pone.0047843. Epub 2012 Oct 29.

iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via

physicochemical properties.

Chen W(1), Lin H, Feng PM, Ding C, Zuo YC, Chou KC.

Author information:

(1)Department of Physics, School of Sciences, Center for Genomics and

Computational Biology, Hebei United University, Tangshan, China.

wchen@gordonlifescience.org

Nucleosome positioning has important roles in key cellular processes. Although

intensive efforts have been made in this area, the rules defining nucleosome

positioning is still elusive and debated. In this study, we carried out a

systematic comparison among the profiles of twelve DNA physicochemical features

between the nucleosomal and linker sequences in the Saccharomyces cerevisiae

genome. We found that nucleosomal sequences have some position-specific

physicochemical features, which can be used for in-depth studying nucleosomes.

Meanwhile, a new predictor, called iNuc-PhysChem, was developed for

identification of nucleosomal sequences by incorporating these physicochemical

properties into a 1788-D (dimensional) feature vector, which was further reduced

to a 884-D vector via the IFS (incremental feature selection) procedure to

optimize the feature set. It was observed by a cross-validation test on a

benchmark dataset that the overall success rate achieved by iNuc-PhysChem was

over 96% in identifying nucleosomal or linker sequences. As a web-server,

iNuc-PhysChem is freely accessible to the public at

http://lin.uestc.edu.cn/server/iNuc-PhysChem. For the convenience of the vast

majority of experimental scientists, a step-by-step guide is provided on how to

use the web-server to get the desired results without the need to follow the

complicated mathematics that were presented just for the integrity in developing

the predictor. Meanwhile, for those who prefer to run predictions in their own

computers, the predictor's code can be easily downloaded from the web-server. It

is anticipated that iNuc-PhysChem may become a useful high throughput tool for

both basic research and drug design.

DOI: 10.1371/journal.pone.0047843

PMCID: PMC3483203

PMID: 23144709 [Indexed for MEDLINE]

1526. PLoS One. 2012;7(10):e48053. doi: 10.1371/journal.pone.0048053. Epub 2012 Oct 24.

SEED servers: high-performance access to the SEED genomes, annotations, and

metabolic models.

Aziz RK(1), Devoid S, Disz T, Edwards RA, Henry CS, Olsen GJ, Olson R, Overbeek

R, Parrello B, Pusch GD, Stevens RL, Vonstein V, Xia F.

Author information:

(1)Computation Institute, University of Chicago, Chicago, Illinois, United States

of America. ramy.aziz@gmail.com

The remarkable advance in sequencing technology and the rising interest in

medical and environmental microbiology, biotechnology, and synthetic biology

resulted in a deluge of published microbial genomes. Yet, genome annotation,

comparison, and modeling remain a major bottleneck to the translation of sequence

information into biological knowledge, hence computational analysis tools are

continuously being developed for rapid genome annotation and interpretation.

Among the earliest, most comprehensive resources for prokaryotic genome analysis,

the SEED project, initiated in 2003 as an integration of genomic data and

analysis tools, now contains >5,000 complete genomes, a constantly updated set of

curated annotations embodied in a large and growing collection of encoded

subsystems, a derived set of protein families, and hundreds of genome-scale

metabolic models. Until recently, however, maintaining current copies of the SEED

code and data at remote locations has been a pressing issue. To allow

high-performance remote access to the SEED database, we developed the SEED

Servers (http://www.theseed.org/servers): four network-based servers intended to

expose the data in the underlying relational database, support basic annotation

services, offer programmatic access to the capabilities of the RAST annotation

server, and provide access to a growing collection of metabolic models that

support flux balance analysis. The SEED servers offer open access to regularly

updated data, the ability to annotate prokaryotic genomes, the ability to create

metabolic reconstructions and detailed models of metabolism, and access to

hundreds of existing metabolic models. This work offers and supports a framework

upon which other groups can build independent research efforts. Large

integrations of genomic data represent one of the major intellectual resources

driving research in biology, and programmatic access to the SEED data will

provide significant utility to a broad collection of potential users.

DOI: 10.1371/journal.pone.0048053

PMCID: PMC3480482

PMID: 23110173 [Indexed for MEDLINE]

1527. PLoS One. 2012;7(10):e47436. doi: 10.1371/journal.pone.0047436. Epub 2012 Oct 15.

Automatic assignment of prokaryotic genes to functional categories using

literature profiling.

Torrieri R(1), Oliveira FS, Oliveira G, Coimbra RS.

Author information:

(1)Center for Excellence in Bioinformatics, FIOCRUZ-Minas, Belo Horizonte,

Brasil.

In the last years, there was an exponential increase in the number of publicly

available genomes. Once finished, most genome projects lack financial support to

review annotations. A few of these gene annotations are based on a combination of

bioinformatics evidence, however, in most cases, annotations are based solely on

sequence similarity to a previously known gene, which was most probably annotated

in the same way. As a result, a large number of predicted genes remain unassigned

to any functional category despite the fact that there is enough evidence in the

literature to predict their function. We developed a classifier trained with

term-frequency vectors automatically disclosed from text corpora of an ensemble

of genes representative of each functional category of the J. Craig Venter

Institute Comprehensive Microbial Resource (JCVI-CMR) ontology. The classifier

achieved up to 84% precision with 68% recall (for confidence≥0.4), F-measure 0.76

(recall and precision equally weighted) in an independent set of 2,220 genes,

from 13 bacterial species, previously classified by JCVI-CMR into unambiguous

categories of its ontology. Finally, the classifier assigned (confidence≥0.7) to

functional categories a total of 5,235 out of the ∼24 thousand genes previously

in categories "Unknown function" or "Unclassified" for which there is literature

in MEDLINE. Two biologists reviewed the literature of 100 of these genes,

randomly picket, and assigned them to the same functional categories predicted by

the automatic classifier. Our results confirmed the hypothesis that it is

possible to confidently assign genes of a real world repository to functional

categories, based exclusively on the automatic profiling of its associated

literature. The LitProf--Gene Classifier web server is accessible at:

www.cebio.org/litprofGC.

DOI: 10.1371/journal.pone.0047436

PMCID: PMC3471813

PMID: 23077617 [Indexed for MEDLINE]

1528. PLoS One. 2012;7(9):e45152. doi: 10.1371/journal.pone.0045152. Epub 2012 Sep 12.

SVMTriP: a method to predict antigenic epitopes using support vector machine to

integrate tri-peptide similarity and propensity.

Yao B(1), Zhang L, Liang S, Zhang C.

Author information:

(1)School of Biological Sciences, Center for Plant Science and Innovation,

University of Nebraska, Lincoln, Nebraska, United States of America.

Identifying protein surface regions preferentially recognizable by antibodies

(antigenic epitopes) is at the heart of new immuno-diagnostic reagent discovery

and vaccine design, and computational methods for antigenic epitope prediction

provide crucial means to serve this purpose. Many linear B-cell epitope

prediction methods were developed, such as BepiPred, ABCPred, AAP, BCPred,

BayesB, BEOracle/BROracle, and BEST, towards this goal. However, effective

immunological research demands more robust performance of the prediction method

than what the current algorithms could provide. In this work, a new method to

predict linear antigenic epitopes is developed; Support Vector Machine has been

utilized by combining the Tri-peptide similarity and Propensity scores (SVMTriP).

Applied to non-redundant B-cell linear epitopes extracted from IEDB, SVMTriP

achieves a sensitivity of 80.1% and a precision of 55.2% with a five-fold

cross-validation. The AUC value is 0.702. The combination of similarity and

propensity of tri-peptide subsequences can improve the prediction performance for

linear B-cell epitopes. Moreover, SVMTriP is capable of recognizing viral

peptides from a human protein sequence background. A web server based on our

method is constructed for public use. The server and all datasets used in the

current study are available at http://sysbio.unl.edu/SVMTriP.

DOI: 10.1371/journal.pone.0045152

PMCID: PMC3440317

PMID: 22984622 [Indexed for MEDLINE]

1529. PLoS One. 2012;7(9):e45125. doi: 10.1371/journal.pone.0045125. Epub 2012 Sep 12.

Insect Innate Immunity Database (IIID): an annotation tool for identifying immune

genes in insect genomes.

Brucker RM(1), Funkhouser LJ, Setia S, Pauly R, Bordenstein SR.

Author information:

(1)Department of Biological Sciences, Vanderbilt University, Nashville,

Tennessee, United States of America. s.bordenstein@vanderbilt.edu

The innate immune system is an ancient component of host defense. Since innate

immunity pathways are well conserved throughout many eukaryotes, immune genes in

model animals can be used to putatively identify homologous genes in newly

sequenced genomes of non-model organisms. With the initiation of the "i5k"

project, which aims to sequence 5,000 insect genomes by 2016, many novel insect

genomes will soon become publicly available, yet few annotation resources are

currently available for insects. Thus, we developed an online tool called the

Insect Innate Immunity Database (IIID) to provide an open access resource for

insect immunity and comparative biology research

(http://www.vanderbilt.edu/IIID). The database provides users with simple

exploratory tools to search the immune repertoires of five insect models

(including Nasonia), spanning three orders, for specific immunity genes or genes

within a particular immunity pathway. As a proof of principle, we used an initial

database with only four insect models to annotate potential immune genes in the

parasitoid wasp genus Nasonia. Results specify 306 putative immune genes in the

genomes of N. vitripennis and its two sister species N. giraulti and N.

longicornis. Of these genes, 146 were not found in previous annotations of

Nasonia immunity genes. Combining these newly identified immune genes with those

in previous annotations, Nasonia possess 489 putative immunity genes, the largest

immune repertoire found in insects to date. While these computational predictions

need to be complemented with functional studies, the IIID database can help

initiate and augment annotations of the immune system in the plethora of insect

genomes that will soon become available.

DOI: 10.1371/journal.pone.0045125

PMCID: PMC3440344

PMID: 22984621 [Indexed for MEDLINE]

1530. PLoS One. 2012;7(9):e42123. doi: 10.1371/journal.pone.0042123. Epub 2012 Sep 10.

A public HTLV-1 molecular epidemiology database for sequence management and data

mining.

Araujo TH(1), Souza-Brito LI, Libin P, Deforche K, Edwards D, de

Albuquerque-Junior AE, Vandamme AM, Galvao-Castro B, Alcantara LC.

Author information:

(1)Gonçalo Moniz Research Center/Oswaldo Cruz Foundation, Salvador, Bahia,

Brazil.

BACKGROUND: It is estimated that 15 to 20 million people are infected with the

human T-cell lymphotropic virus type 1 (HTLV-1). At present, there are more than

2,000 unique HTLV-1 isolate sequences published. A central database to aggregate

sequence information from a range of epidemiological aspects including HTLV-1

infections, pathogenesis, origins, and evolutionary dynamics would be useful to

scientists and physicians worldwide. Described here, we have developed a database

that collects and annotates sequence data and can be accessed through a

user-friendly search interface. The HTLV-1 Molecular Epidemiology Database

website is available at http://htlv1db.bahia.fiocruz.br/.

METHODOLOGY/PRINCIPAL FINDINGS: All data was obtained from publications available

at GenBank or through contact with the authors. The database was developed using

Apache Webserver 2.1.6 and SGBD MySQL. The webpage interfaces were developed in

HTML and sever-side scripting written in PHP. The HTLV-1 Molecular Epidemiology

Database is hosted on the Gonçalo Moniz/FIOCRUZ Research Center server. There are

currently 2,457 registered sequences with 2,024 (82.37%) of those sequences

representing unique isolates. Of these sequences, 803 (39.67%) contain

information about clinical status (TSP/HAM, 17.19%; ATL, 7.41%; asymptomatic,

12.89%; other diseases, 2.17%; and no information, 60.32%). Further, 7.26% of

sequences contain information on patient gender while 5.23% of sequences provide

the age of the patient.

CONCLUSIONS/SIGNIFICANCE: The HTLV-1 Molecular Epidemiology Database retrieves

and stores annotated HTLV-1 proviral sequences from clinical, epidemiological,

and geographical studies. The collected sequences and related information are now

accessible on a publically available and user-friendly website. This open-access

database will support clinical research and vaccine development related to viral

genotype.

DOI: 10.1371/journal.pone.0042123

PMCID: PMC3438164

PMID: 22970114 [Indexed for MEDLINE]

1531. PLoS One. 2012;7(8):e43575. doi: 10.1371/journal.pone.0043575. Epub 2012 Aug 21.

Computational prediction of conformational B-cell epitopes from antigen primary

structures by ensemble learning.

Zhang W(1), Niu Y, Xiong Y, Zhao M, Yu R, Liu J.

Author information:

(1)School of Computer, Wuhan University, Wuhan, China. zhangwen@whu.edu.cn

MOTIVATION: The conformational B-cell epitopes are the specific sites on the

antigens that have immune functions. The identification of conformational B-cell

epitopes is of great importance to immunologists for facilitating the design of

peptide-based vaccines. As an attempt to narrow the search for experimental

validation, various computational models have been developed for the epitope

prediction by using antigen structures. However, the application of these models

is undermined by the limited number of available antigen structures. In contrast

to the most of available structure-based methods, we here attempt to accurately

predict conformational B-cell epitopes from antigen sequences.

METHODS: In this paper, we explore various sequence-derived features, which have

been observed to be associated with the location of epitopes or ever used in the

similar tasks. These features are evaluated and ranked by their discriminative

performance on the benchmark datasets. From the perspective of information

science, the combination of various features can usually lead to better results

than the individual features. In order to build the robust model, we adopt the

ensemble learning approach to incorporate various features, and develop the

ensemble model to predict conformational epitopes from antigen sequences.

RESULTS: Evaluated by the leave-one-out cross validation, the proposed method

gives out the mean AUC scores of 0.687 and 0.651 on two datasets respectively

compiled from the bound structures and unbound structures. When compared with

publicly available servers by using the independent dataset, our method yields

better or comparable performance. The results demonstrate the proposed method is

useful for the sequence-based conformational epitope prediction.

AVAILABILITY: The web server and datasets are freely available at

http://bcell.whu.edu.cn.

DOI: 10.1371/journal.pone.0043575

PMCID: PMC3424238

PMID: 22927994 [Indexed for MEDLINE]

1532. PLoS One. 2012;7(8):e42390. doi: 10.1371/journal.pone.0042390. Epub 2012 Aug 1.

miRSystem: an integrated system for characterizing enriched functions and

pathways of microRNA targets.

Lu TP(1), Lee CY, Tsai MH, Chiu YC, Hsiao CK, Lai LC, Chuang EY.

Author information:

(1)YongLin Biomedical Engineering Center, National Taiwan University, Taipei,

Taiwan.

BACKGROUND: Many prediction tools for microRNA (miRNA) targets have been

developed, but inconsistent predictions were observed across multiple algorithms,

which can make further analysis difficult. Moreover, the nomenclature of human

miRNAs changes rapidly. To address these issues, we developed a web-based system,

miRSystem, for converting queried miRNAs to the latest annotation and predicting

the function of miRNA by integrating miRNA target gene prediction and

function/pathway analyses.

RESULTS: First, queried miRNA IDs were converted to the latest annotated version

to prevent potential conflicts resulting from multiple aliases. Next, by

combining seven algorithms and two validated databases, potential gene targets of

miRNAs and their functions were predicted based on the consistency across

independent algorithms and observed/expected ratios. Lastly, five pathway

databases were included to characterize the enriched pathways of target genes

through bootstrap approaches. Based on the enriched pathways of target genes, the

functions of queried miRNAs could be predicted.

CONCLUSIONS: MiRSystem is a user-friendly tool for predicting the target genes

and their associated pathways for many miRNAs simultaneously. The web server and

the documentation are freely available at http://mirsystem.cgm.ntu.edu.tw/.

DOI: 10.1371/journal.pone.0042390

PMCID: PMC3411648

PMID: 22870325 [Indexed for MEDLINE]

1533. PLoS One. 2012;7(8):e41827. doi: 10.1371/journal.pone.0041827. Epub 2012 Aug 3.

NEXCADE: perturbation analysis for complex networks.

Yadav G(1), Babu S.

Author information:

(1)Computational Biology Laboratory, National Institute of Plant Genome Research,

New Delhi, India. gy@nipgr.res.in

Recent advances in network theory have led to considerable progress in our

understanding of complex real world systems and their behavior in response to

external threats or fluctuations. Much of this research has been invigorated by

demonstration of the 'robust, yet fragile' nature of cellular and large-scale

systems transcending biology, sociology, and ecology, through application of the

network theory to diverse interactions observed in nature such as

plant-pollinator, seed-dispersal agent and host-parasite relationships. In this

work, we report the development of NEXCADE, an automated and interactive program

for inducing disturbances into complex systems defined by networks, focusing on

the changes in global network topology and connectivity as a function of the

perturbation. NEXCADE uses a graph theoretical approach to simulate perturbations

in a user-defined manner, singly, in clusters, or sequentially. To demonstrate

the promise it holds for broader adoption by the research community, we provide

pre-simulated examples from diverse real-world networks including eukaryotic

protein-protein interaction networks, fungal biochemical networks, a variety of

ecological food webs in nature as well as social networks. NEXCADE not only

enables network visualization at every step of the targeted attacks, but also

allows risk assessment, i.e. identification of nodes critical for the robustness

of the system of interest, in order to devise and implement context-based

strategies for restructuring a network, or to achieve resilience against link or

node failures. Source code and license for the software, designed to work on a

Linux-based operating system (OS) can be downloaded at

http://www.nipgr.res.in/nexcade\_download.html. In addition, we have developed

NEXCADE as an OS-independent online web server freely available to the scientific

community without any login requirement at http://www.nipgr.res.in/nexcade.html.

DOI: 10.1371/journal.pone.0041827

PMCID: PMC3411682

PMID: 22870252 [Indexed for MEDLINE]

1534. PLoS One. 2012;7(7):e41202. doi: 10.1371/journal.pone.0041202. Epub 2012 Jul 24.

Prediction and analysis of the protein interactome in Pseudomonas aeruginosa to

enable network-based drug target selection.

Zhang M(1), Su S, Bhatnagar RK, Hassett DJ, Lu LJ.

Author information:

(1)Division of Biomedical Informatics, Cincinnati Children's Hospital Research

Foundation, Cincinnati, Ohio, United States of America.

Pseudomonas aeruginosa (PA) is a ubiquitous opportunistic pathogen that is

capable of causing highly problematic, chronic infections in cystic fibrosis and

chronic obstructive pulmonary disease patients. With the increased prevalence of

multi-drug resistant PA, the conventional "one gene, one drug, one disease"

paradigm is losing effectiveness. Network pharmacology, on the other hand, may

hold the promise of discovering new drug targets to treat a variety of PA

infections. However, given the urgent need for novel drug target discovery, a PA

protein-protein interaction (PPI) network of high accuracy and coverage, has not

yet been constructed. In this study, we predicted a genome-scale PPI network of

PA by integrating various genomic features of PA proteins/genes by a machine

learning-based approach. A total of 54,107 interactions covering 4,181 proteins

in PA were predicted. A high-confidence network combining predicted

high-confidence interactions, a reference set and verified interactions that

consist of 3,343 proteins and 19,416 potential interactions was further assembled

and analyzed. The predicted interactome network from this study is the first

large-scale PPI network in PA with significant coverage and high accuracy.

Subsequent analysis, including validations based on existing small-scale PPI data

and the network structure comparison with other model organisms, shows the

validity of the predicted PPI network. Potential drug targets were identified and

prioritized based on their essentiality and topological importance in the

high-confidence network. Host-pathogen protein interactions between human and PA

were further extracted and analyzed. In addition, case studies were performed on

protein interactions regarding anti-sigma factor MucA, negative periplasmic

alginate regulator MucB, and the transcriptional regulator RhlR. A web server to

access the predicted PPI dataset is available at

http://research.cchmc.org/PPIdatabase/.

DOI: 10.1371/journal.pone.0041202

PMCID: PMC3404098

PMID: 22848443 [Indexed for MEDLINE]

1535. PLoS One. 2012;7(7):e40155. doi: 10.1371/journal.pone.0040155. Epub 2012 Jul 9.

GlycoPP: a webserver for prediction of N- and O-glycosites in prokaryotic protein

sequences.

Chauhan JS(1), Bhat AH, Raghava GP, Rao A.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Council of

Scientific and Industrial Research, Chandigarh, India.

Glycosylation is one of the most abundant post-translational modifications (PTMs)

required for various structure/function modulations of proteins in a living cell.

Although elucidated recently in prokaryotes, this type of PTM is present across

all three domains of life. In prokaryotes, two types of protein glycan linkages

are more widespread namely, N- linked, where a glycan moiety is attached to the

amide group of Asn, and O- linked, where a glycan moiety is attached to the

hydroxyl group of Ser/Thr/Tyr. For their biologically ubiquitous nature,

significance, and technology applications, the study of prokaryotic glycoproteins

is a fast emerging area of research. Here we describe new Support Vector Machine

(SVM) based algorithms (models) developed for predicting glycosylated-residues

(glycosites) with high accuracy in prokaryotic protein sequences. The models are

based on binary profile of patterns, composition profile of patterns, and

position-specific scoring matrix profile of patterns as training features. The

study employ an extensive dataset of 107 N-linked and 116 O-linked glycosites

extracted from 59 experimentally characterized glycoproteins of prokaryotes. This

dataset includes validated N-glycosites from phyla Crenarchaeota, Euryarchaeota

(domain Archaea), Proteobacteria (domain Bacteria) and validated O-glycosites

from phyla Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria (domain

Bacteria). In view of the current understanding that glycosylation occurs on

folded proteins in bacteria, hybrid models have been developed using information

on predicted secondary structures and accessible surface area in various

combinations with training features. Using these models, N-glycosites and

O-glycosites could be predicted with an accuracy of 82.71% (MCC 0.65) and 73.71%

(MCC 0.48), respectively. An evaluation of the best performing models with 28

independent prokaryotic glycoproteins confirms the suitability of these models in

predicting N- and O-glycosites in potential glycoproteins from aforementioned

organisms, with reasonably high confidence. A web server GlycoPP, implementing

these models is available freely at http:/www.imtech.res.in/raghava/glycopp/.

DOI: 10.1371/journal.pone.0040155

PMCID: PMC3392279

PMID: 22808107 [Indexed for MEDLINE]

1536. PLoS One. 2012;7(6):e38979. doi: 10.1371/journal.pone.0038979. Epub 2012 Jun 18.

A global characterization and identification of multifunctional enzymes.

Cheng XY(1), Huang WJ, Hu SC, Zhang HL, Wang H, Zhang JX, Lin HH, Chen YZ, Zou Q,

Ji ZL.

Author information:

(1)State Key Laboratory of Stress Cell Biology, School of Life Sciences, Xiamen

University, Xiamen, Fujian, People's Republic of China.

Multi-functional enzymes are enzymes that perform multiple physiological

functions. Characterization and identification of multi-functional enzymes are

critical for communication and cooperation between different functions and

pathways within a complex cellular system or between cells. In present study, we

collected literature-reported 6,799 multi-functional enzymes and systematically

characterized them in structural, functional, and evolutionary aspects. It was

found that four physiochemical properties, that is, charge, polarizability,

hydrophobicity, and solvent accessibility, are important for characterization of

multi-functional enzymes. Accordingly, a combinational model of support vector

machine and random forest model was constructed, based on which 6,956 potential

novel multi-functional enzymes were successfully identified from the ENZYME

database. Moreover, it was observed that multi-functional enzymes are non-evenly

distributed in species, and that Bacteria have relatively more multi-functional

enzymes than Archaebacteria and Eukaryota. Comparative analysis indicated that

the multi-functional enzymes experienced a fluctuation of gene gain and loss

during the evolution from S. cerevisiae to H. sapiens. Further pathway analyses

indicated that a majority of multi-functional enzymes were well preserved in

catalyzing several essential cellular processes, for example, metabolisms of

carbohydrates, nucleotides, and amino acids. What's more, a database of known

multi-functional enzymes and a server for novel multi-functional enzyme

prediction were also constructed for free access at

http://bioinf.xmu.edu.cn/databases/MFEs/index.htm.

DOI: 10.1371/journal.pone.0038979

PMCID: PMC3377604

PMID: 22723914 [Indexed for MEDLINE]

1537. PLoS One. 2012;7(5):e37869. doi: 10.1371/journal.pone.0037869. Epub 2012 May 25.

PepMapper: a collaborative web tool for mapping epitopes from affinity-selected

peptides.

Chen W(1), Guo WW, Huang Y, Ma Z.

Author information:

(1)School of Computer Science and Information Technology, Northeast Normal

University, Changchun, China.

Epitope mapping from affinity-selected peptides has become popular in epitope

prediction, and correspondingly many Web-based tools have been developed in

recent years. However, the performance of these tools varies in different

circumstances. To address this problem, we employed an ensemble approach to

incorporate two popular Web tools, MimoPro and Pep-3D-Search, together for taking

advantages offered by both methods so as to give users more options for their

specific purposes of epitope-peptide mapping. The combined operation of Union

finds as many associated peptides as possible from both methods, which increases

sensitivity in finding potential epitopic regions on a given antigen surface. The

combined operation of Intersection achieves to some extent the mutual

verification by the two methods and hence increases the likelihood of locating

the genuine epitopic region on a given antigen in relation to the interacting

peptides. The Consistency between Intersection and Union is an indirect

sufficient condition to assess the likelihood of successful peptide-epitope

mapping. On average from 27 tests, the combined operations of PepMapper

outperformed either MimoPro or Pep-3D-Search alone. Therefore, PepMapper is

another multipurpose mapping tool for epitope prediction from affinity-selected

peptides. The Web server can be freely accessed at:

http://informatics.nenu.edu.cn/PepMapper/

DOI: 10.1371/journal.pone.0037869

PMCID: PMC3360666

PMID: 22701536 [Indexed for MEDLINE]

1538. PLoS One. 2012;7(5):e36317. doi: 10.1371/journal.pone.0036317. Epub 2012 May 22.

A multi-label predictor for identifying the subcellular locations of singleplex

and multiplex eukaryotic proteins.

Wang X(1), Li GZ.

Author information:

(1)The MOE Key Laboratory of Embedded System and Service Computing, Department of

Control Science and Engineering, Tongji University, Shanghai, China.

Subcellular locations of proteins are important functional attributes. An

effective and efficient subcellular localization predictor is necessary for

rapidly and reliably annotating subcellular locations of proteins. Most of

existing subcellular localization methods are only used to deal with

single-location proteins. Actually, proteins may simultaneously exist at, or move

between, two or more different subcellular locations. To better reflect

characteristics of multiplex proteins, it is highly desired to develop new

methods for dealing with them. In this paper, a new predictor, called

Euk-ECC-mPLoc, by introducing a powerful multi-label learning approach which

exploits correlations between subcellular locations and hybridizing gene ontology

with dipeptide composition information, has been developed that can be used to

deal with systems containing both singleplex and multiplex eukaryotic proteins.

It can be utilized to identify eukaryotic proteins among the following 22

locations: (1) acrosome, (2) cell membrane, (3) cell wall, (4) centrosome, (5)

chloroplast, (6) cyanelle, (7) cytoplasm, (8) cytoskeleton, (9) endoplasmic

reticulum, (10) endosome, (11) extracellular, (12) Golgi apparatus, (13)

hydrogenosome, (14) lysosome, (15) melanosome, (16) microsome, (17)

mitochondrion, (18) nucleus, (19) peroxisome, (20) spindle pole body, (21)

synapse, and (22) vacuole. Experimental results on a stringent benchmark dataset

of eukaryotic proteins by jackknife cross validation test show that the average

success rate and overall success rate obtained by Euk-ECC-mPLoc were 69.70% and

81.54%, respectively, indicating that our approach is quite promising.

Particularly, the success rates achieved by Euk-ECC-mPLoc for small subsets were

remarkably improved, indicating that it holds a high potential for simulating the

development of the area. As a user-friendly web-server, Euk-ECC-mPLoc is freely

accessible to the public at the website

http://levis.tongji.edu.cn:8080/bioinfo/Euk-ECC-mPLoc/. We believe that

Euk-ECC-mPLoc may become a useful high-throughput tool, or at least play a

complementary role to the existing predictors in identifying subcellular

locations of eukaryotic proteins.

DOI: 10.1371/journal.pone.0036317

PMCID: PMC3358325

PMID: 22629314 [Indexed for MEDLINE]

1539. PLoS One. 2012;7(5):e35146. doi: 10.1371/journal.pone.0035146. Epub 2012 May 4.

DNA barcode goes two-dimensions: DNA QR code web server.

Liu C(1), Shi L, Xu X, Li H, Xing H, Liang D, Jiang K, Pang X, Song J, Chen S.

Author information:

(1)Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences,

Peking Union Medical College, Beijing, People's Republic of China.

cliu@implad.ac.cn

The DNA barcoding technology uses a standard region of DNA sequence for species

identification and discovery. At present, "DNA barcode" actually refers to DNA

sequences, which are not amenable to information storage, recognition, and

retrieval. Our aim is to identify the best symbology that can represent DNA

barcode sequences in practical applications. A comprehensive set of sequences for

five DNA barcode markers ITS2, rbcL, matK, psbA-trnH, and CO1 was used as the

test data. Fifty-three different types of one-dimensional and ten two-dimensional

barcode symbologies were compared based on different criteria, such as coding

capacity, compression efficiency, and error detection ability. The quick response

(QR) code was found to have the largest coding capacity and relatively high

compression ratio. To facilitate the further usage of QR code-based DNA barcodes,

a web server was developed and is accessible at http://qrfordna.dnsalias.org. The

web server allows users to retrieve the QR code for a species of interests,

convert a DNA sequence to and from a QR code, and perform species identification

based on local and global sequence similarities. In summary, the first

comprehensive evaluation of various barcode symbologies has been carried out. The

QR code has been found to be the most appropriate symbology for DNA barcode

sequences. A web server has also been constructed to allow biologists to utilize

QR codes in practical DNA barcoding applications.

DOI: 10.1371/journal.pone.0035146

PMCID: PMC3344831

PMID: 22574113 [Indexed for MEDLINE]

1540. PLoS One. 2012;7(4):e36202. doi: 10.1371/journal.pone.0036202. Epub 2012 Apr 27.

Linking proteins to signaling pathways for experiment design and evaluation.

Farkas IJ(1), Szántó-Várnagy A, Korcsmáros T.

Author information:

(1)Statistical and Biological Physics Research Group, Hungarian Academy of

Sciences, Budapest, Hungary. fij@elte.hu

Biomedical experimental work often focuses on altering the functions of selected

proteins. These changes can hit signaling pathways, and can therefore

unexpectedly and non-specifically affect cellular processes. We propose

PathwayLinker, an online tool that can provide a first estimate of the possible

signaling effects of such changes, e.g., drug or microRNA treatments.

PathwayLinker minimizes the users' efforts by integrating protein-protein

interaction and signaling pathway data from several sources with statistical

significance tests and clear visualization. We demonstrate through three case

studies that the developed tool can point out unexpected signaling bias in normal

laboratory experiments and identify likely novel signaling proteins among the

interactors of known drug targets. In our first case study we show that knockdown

of the Caenorhabditis elegans gene cdc-25.1 (meant to avoid progeny) may globally

affect the signaling system and unexpectedly bias experiments. In the second case

study we evaluate the loss-of-function phenotypes of a less known C. elegans gene

to predict its function. In the third case study we analyze GJA1, an anti-cancer

drug target protein in human, and predict for this protein novel signaling

pathway memberships, which may be sources of side effects. Compared to similar

services, a major advantage of PathwayLinker is that it drastically reduces the

necessary amount of manual literature searches and can be used without a

computational background. PathwayLinker is available at http://PathwayLinker.org.

Detailed documentation and source code are available at the website.

DOI: 10.1371/journal.pone.0036202

PMCID: PMC3338605

PMID: 22558382 [Indexed for MEDLINE]

1541. PLoS One. 2012;7(4):e34030. doi: 10.1371/journal.pone.0034030. Epub 2012 Apr 4.

Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin.

Sharma VK(1), Kumar N, Prakash T, Taylor TD.

Author information:

(1)Laboratory for MetaSystems Research, Quantitative Biology Center, RIKEN,

Yokohama, Kanagawa, Japan.

Taxonomic assignment of sequence reads is a challenging task in metagenomic data

analysis, for which the present methods mainly use either composition- or

homology-based approaches. Though the homology-based methods are more sensitive

and accurate, they suffer primarily due to the time needed to generate the Blast

alignments. We developed the MetaBin program and web server for better

homology-based taxonomic assignments using an ORF-based approach. By implementing

Blat as the faster alignment method in place of Blastx, the analysis time has

been reduced by severalfold. It is benchmarked using both simulated and real

metagenomic datasets, and can be used for both single and paired-end sequence

reads of varying lengths (≥45 bp). To our knowledge, MetaBin is the only

available program that can be used for the taxonomic binning of short reads (<100

bp) with high accuracy and high sensitivity using a homology-based approach. The

MetaBin web server can be used to carry out the taxonomic analysis, by either

submitting reads or Blastx output. It provides several options including

construction of taxonomic trees, creation of a composition chart, functional

analysis using COGs, and comparative analysis of multiple metagenomic datasets.

MetaBin web server and a standalone version for high-throughput analysis are

available freely at http://metabin.riken.jp/.

DOI: 10.1371/journal.pone.0034030

PMCID: PMC3319535

PMID: 22496776 [Indexed for MEDLINE]

1542. PLoS One. 2012;7(3):e32833. doi: 10.1371/journal.pone.0032833. Epub 2012 Mar 14.

UniDrug-target: a computational tool to identify unique drug targets in

pathogenic bacteria.

Chanumolu SK(1), Rout C, Chauhan RS.

Author information:

(1)Department of Biotechnology and Bioinformatics, Jaypee University of

Information Technology, Waknaghat, Solan, Himachal Pradesh, India.

BACKGROUND: Targeting conserved proteins of bacteria through antibacterial

medications has resulted in both the development of resistant strains and changes

to human health by destroying beneficial microbes which eventually become

breeding grounds for the evolution of resistances. Despite the availability of

more than 800 genomes sequences, 430 pathways, 4743 enzymes, 9257 metabolic

reactions and protein (three-dimensional) 3D structures in bacteria, no

pathogen-specific computational drug target identification tool has been

developed.

METHODS: A web server, UniDrug-Target, which combines bacterial biological

information and computational methods to stringently identify pathogen-specific

proteins as drug targets, has been designed. Besides predicting pathogen-specific

proteins essentiality, chokepoint property, etc., three new algorithms were

developed and implemented by using protein sequences, domains, structures, and

metabolic reactions for construction of partial metabolic networks (PMNs),

determination of conservation in critical residues, and variation analysis of

residues forming similar cavities in proteins sequences. First, PMNs are

constructed to determine the extent of disturbances in metabolite production by

targeting a protein as drug target. Conservation of pathogen-specific protein's

critical residues involved in cavity formation and biological function determined

at domain-level with low-matching sequences. Last, variation analysis of residues

forming similar cavities in proteins sequences from pathogenic versus

non-pathogenic bacteria and humans is performed.

RESULTS: The server is capable of predicting drug targets for any sequenced

pathogenic bacteria having fasta sequences and annotated information. The utility

of UniDrug-Target server was demonstrated for Mycobacterium tuberculosis (H37Rv).

The UniDrug-Target identified 265 mycobacteria pathogen-specific proteins,

including 17 essential proteins which can be potential drug targets.

CONCLUSIONS/SIGNIFICANCE: UniDrug-Target is expected to accelerate

pathogen-specific drug targets identification which will increase their success

and durability as drugs developed against them have less chance to develop

resistances and adverse impact on environment. The server is freely available at

http://117.211.115.67/UDT/main.html. The standalone application (source codes) is

available at http://www.bioinformatics.org/ftp/pub/bioinfojuit/UDT.rar.

DOI: 10.1371/journal.pone.0032833

PMCID: PMC3303792

PMID: 22431985 [Indexed for MEDLINE]

1543. PLoS One. 2012;7(3):e31362. doi: 10.1371/journal.pone.0031362. Epub 2012 Mar 7.

Phylo: a citizen science approach for improving multiple sequence alignment.

Kawrykow A(1), Roumanis G, Kam A, Kwak D, Leung C, Wu C, Zarour E; Phylo players,

Sarmenta L, Blanchette M, Waldispühl J.

Author information:

(1)School of Computer Science and McGill Centre for Bioinformatics, McGill

University, Montreal, Quebec, Canada.

BACKGROUND: Comparative genomics, or the study of the relationships of genome

structure and function across different species, offers a powerful tool for

studying evolution, annotating genomes, and understanding the causes of various

genetic disorders. However, aligning multiple sequences of DNA, an essential

intermediate step for most types of analyses, is a difficult computational task.

In parallel, citizen science, an approach that takes advantage of the fact that

the human brain is exquisitely tuned to solving specific types of problems, is

becoming increasingly popular. There, instances of hard computational problems

are dispatched to a crowd of non-expert human game players and solutions are sent

back to a central server.

METHODOLOGY/PRINCIPAL FINDINGS: We introduce Phylo, a human-based computing

framework applying "crowd sourcing" techniques to solve the Multiple Sequence

Alignment (MSA) problem. The key idea of Phylo is to convert the MSA problem into

a casual game that can be played by ordinary web users with a minimal prior

knowledge of the biological context. We applied this strategy to improve the

alignment of the promoters of disease-related genes from up to 44 vertebrate

species. Since the launch in November 2010, we received more than 350,000

solutions submitted from more than 12,000 registered users. Our results show that

solutions submitted contributed to improving the accuracy of up to 70% of the

alignment blocks considered.

CONCLUSIONS/SIGNIFICANCE: We demonstrate that, combined with classical

algorithms, crowd computing techniques can be successfully used to help improving

the accuracy of MSA. More importantly, we show that an NP-hard computational

problem can be embedded in casual game that can be easily played by people

without significant scientific training. This suggests that citizen science

approaches can be used to exploit the billions of "human-brain peta-flops" of

computation that are spent every day playing games. Phylo is available at:

http://phylo.cs.mcgill.ca.

DOI: 10.1371/journal.pone.0031362

PMCID: PMC3296692

PMID: 22412834 [Indexed for MEDLINE]

1544. PLoS One. 2012;7(2):e30483. doi: 10.1371/journal.pone.0030483. Epub 2012 Feb 23.

TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700

HLA-DR molecules.

Zhang L(1), Chen Y, Wong HS, Zhou S, Mamitsuka H, Zhu S.

Author information:

(1)School of Computer Science and Shanghai Key Lab of Intelligent Information

Processing, Fudan University, Shanghai, China.

MOTIVATION: Accurate identification of peptides binding to specific Major

Histocompatibility Complex Class II (MHC-II) molecules is of great importance for

elucidating the underlying mechanism of immune recognition, as well as for

developing effective epitope-based vaccines and promising immunotherapies for

many severe diseases. Due to extreme polymorphism of MHC-II alleles and the high

cost of biochemical experiments, the development of computational methods for

accurate prediction of binding peptides of MHC-II molecules, particularly for the

ones with few or no experimental data, has become a topic of increasing interest.

TEPITOPE is a well-used computational approach because of its good

interpretability and relatively high performance. However, TEPITOPE can be

applied to only 51 out of over 700 known HLA DR molecules.

METHOD: We have developed a new method, called TEPITOPEpan, by extrapolating from

the binding specificities of HLA DR molecules characterized by TEPITOPE to those

uncharacterized. First, each HLA-DR binding pocket is represented by amino acid

residues that have close contact with the corresponding peptide binding core

residues. Then the pocket similarity between two HLA-DR molecules is calculated

as the sequence similarity of the residues. Finally, for an uncharacterized

HLA-DR molecule, the binding specificity of each pocket is computed as a weighted

average in pocket binding specificities over HLA-DR molecules characterized by

TEPITOPE.

RESULT: The performance of TEPITOPEpan has been extensively evaluated using

various data sets from different viewpoints: predicting MHC binding peptides,

identifying HLA ligands and T-cell epitopes and recognizing binding cores. Among

the four state-of-the-art competing pan-specific methods, for predicting binding

specificities of unknown HLA-DR molecules, TEPITOPEpan was roughly the second

best method next to NETMHCIIpan-2.0. Additionally, TEPITOPEpan achieved the best

performance in recognizing binding cores. We further analyzed the motifs detected

by TEPITOPEpan, examining the corresponding literature of immunology. Its online

server and PSSMs therein are available at

http://www.biokdd.fudan.edu.cn/Service/TEPITOPEpan/.

DOI: 10.1371/journal.pone.0030483

PMCID: PMC3285624

PMID: 22383964 [Indexed for MEDLINE]

1545. PLoS One. 2012;7(2):e30869. doi: 10.1371/journal.pone.0030869. Epub 2012 Feb 21.

iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and

their subfamilies via physical-chemical property matrix.

Xiao X(1), Wang P, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China.

xiaoxuan0326@yahoo.com.cn

Nuclear receptors (NRs) form a family of ligand-activated transcription factors

that regulate a wide variety of biological processes, such as homeostasis,

reproduction, development, and metabolism. Human genome contains 48 genes

encoding NRs. These receptors have become one of the most important targets for

therapeutic drug development. According to their different action mechanisms or

functions, NRs have been classified into seven subfamilies. With the avalanche of

protein sequences generated in the postgenomic age, we are facing the following

challenging problems. Given an uncharacterized protein sequence, how can we

identify whether it is a nuclear receptor? If it is, what subfamily it belongs

to? To address these problems, we developed a predictor called iNR-PhysChem in

which the protein samples were expressed by a novel mode of pseudo amino acid

composition (PseAAC) whose components were derived from a physical-chemical

matrix via a series of auto-covariance and cross-covariance transformations. It

was observed that the overall success rate achieved by iNR-PhysChem was over 98%

in identifying NRs or non-NRs, and over 92% in identifying NRs among the

following seven subfamilies: NR1--thyroid hormone like, NR2--HNF4-like,

NR3--estrogen like, NR4--nerve growth factor IB-like, NR5--fushi tarazu-F1 like,

NR6--germ cell nuclear factor like, and NR0--knirps like. These rates were

derived by the jackknife tests on a stringent benchmark dataset in which none of

protein sequences included has ≥60% pairwise sequence identity to any other in a

same subset. As a user-friendly web-server, iNR-PhysChem is freely accessible to

the public at either http://www.jci-bioinfo.cn/iNR-PhysChem or

http://icpr.jci.edu.cn/bioinfo/iNR-PhysChem. Also a step-by-step guide is

provided on how to use the web-server to get the desired results without the need

to follow the complicated mathematics involved in developing the predictor. It is

anticipated that iNR-PhysChem may become a useful high throughput tool for both

basic research and drug design.

DOI: 10.1371/journal.pone.0030869

PMCID: PMC3283608

PMID: 22363503 [Indexed for MEDLINE]

1546. PLoS One. 2012;7(1):e30737. doi: 10.1371/journal.pone.0030737. Epub 2012 Jan 27.

Catalog of microRNA seed polymorphisms in vertebrates.

Zorc M(1), Skok DJ, Godnic I, Calin GA, Horvat S, Jiang Z, Dovc P, Kunej T.

Author information:

(1)Department of Animal Science, Biotechnical Faculty, University of Ljubljana,

Domzale, Slovenia.

MicroRNAs (miRNAs) are a class of non-coding RNA that plays an important role in

posttranscriptional regulation of mRNA. Evidence has shown that miRNA gene

variability might interfere with its function resulting in phenotypic variation

and disease susceptibility. A major role in miRNA target recognition is ascribed

to complementarity with the miRNA seed region that can be affected by

polymorphisms. In the present study, we developed an online tool for the

detection of miRNA polymorphisms (miRNA SNiPer) in vertebrates

(http://www.integratomics-time.com/miRNA-SNiPer) and generated a catalog of miRNA

seed region polymorphisms (miR-seed-SNPs) consisting of 149 SNPs in six species.

Although a majority of detected polymorphisms were due to point mutations, two

consecutive nucleotide substitutions (double nucleotide polymorphisms, DNPs) were

also identified in nine miRNAs. We determined that miR-SNPs are frequently

located within the quantitative trait loci (QTL), chromosome fragile sites, and

cancer susceptibility loci, indicating their potential role in the genetic

control of various complex traits. To test this further, we performed an

association analysis between the mmu-miR-717 seed SNP rs30372501, which is

polymorphic in a large number of standard inbred strains, and all phenotypic

traits in these strains deposited in the Mouse Phenome Database. Analysis showed

a significant association between the mmu-miR-717 seed SNP and a diverse array of

traits including behavior, blood-clinical chemistry, body weight size and growth,

and immune system suggesting that seed SNPs can indeed have major pleiotropic

effects. The bioinformatics analyses, data and tools developed in the present

study can serve researchers as a starting point in testing more targeted

hypotheses and designing experiments using optimal species or strains for further

mechanistic studies.

DOI: 10.1371/journal.pone.0030737

PMCID: PMC3267754

PMID: 22303453 [Indexed for MEDLINE]

1547. PLoS One. 2012;7(1):e30938. doi: 10.1371/journal.pone.0030938. Epub 2012 Jan 26.

SPPS: a sequence-based method for predicting probability of protein-protein

interaction partners.

Liu X(1), Liu B, Huang Z, Shi T, Chen Y, Zhang J.

Author information:

(1)Department of Pathophysiology, Key Laboratory of Cell Differentiation and

Apoptosis of Chinese Ministry of Education, School of Medicine, Shanghai Jiao

Tong University, Shanghai, China.

BACKGROUND: The molecular network sustained by different types of interactions

among proteins is widely manifested as the fundamental driving force of cellular

operations. Many biological functions are determined by the crosstalk between

proteins rather than by the characteristics of their individual components. Thus,

the searches for protein partners in global networks are imperative when

attempting to address the principles of biology.

RESULTS: We have developed a web-based tool "Sequence-based Protein Partners

Search" (SPPS) to explore interacting partners of proteins, by searching over a

large repertoire of proteins across many species. SPPS provides a database

containing more than 60,000 protein sequences with annotations and a

protein-partner search engine in two modes (Single Query and Multiple Query). Two

interacting proteins of human FBXO6 protein have been found using the service in

the study. In addition, users can refine potential protein partner hits by using

annotations and possible interactive network in the SPPS web server.

CONCLUSIONS: SPPS provides a new type of tool to facilitate the identification of

direct or indirect protein partners which may guide scientists on the

investigation of new signaling pathways. The SPPS server is available to the

public at http://mdl.shsmu.edu.cn/SPPS/.

DOI: 10.1371/journal.pone.0030938

PMCID: PMC3266917

PMID: 22292078 [Indexed for MEDLINE]

1548. Protein Cell. 2012 Jan;3(1):38-43. doi: 10.1007/s13238-011-1130-2. Epub 2011 Dec

19.

SySAP: a system-level predictor of deleterious single amino acid polymorphisms.

Huang T(1), Wang C, Zhang G, Xie L, Li Y.

Author information:

(1)Key Laboratory of Systems Biology, Shanghai Institutes for Biological

Sciences, Chinese Academy of Sciences, Shanghai, China.

Single amino acid polymorphisms (SAPs), also known as non-synonymous single

nucleotide polymorphisms (nsSNPs), are responsible for most of human genetic

diseases. Discriminate the deleterious SAPs from neutral ones can help identify

the disease genes and understand the mechanism of diseases. In this work, a

method of deleterious SAP prediction at system level was established. Unlike most

existing methods, our method not only considers the sequence and structure

information, but also the network information. The integration of network

information can improve the performance of deleterious SAP prediction. To make

our method available to the public, we developed SySAP (a System-level predictor

of deleterious Single Amino acid Polymorphisms), an easy-to-use and high accurate

web server. SySAP is freely available at http://www.biosino.org/ SySAP/ and

http://lifecenter.sgst.cn/SySAP/.

DOI: 10.1007/s13238-011-1130-2

PMCID: PMC4875213

PMID: 22183811 [Indexed for MEDLINE]

1549. Protein Pept Lett. 2012 Jan;19(1):4-14.

iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization

of singleplex and multiplex Gram-positive bacterial proteins.

Wu ZC(1), Xiao X, Chou KC.

Author information:

(1)Computer Department, Jing- De-Zhen Ceramic Institute, Jing-De-Zhen, China.

By introducing the "multi-layer scale", as well as hybridizing the information of

gene ontology and the sequential evolution information, a novel predictor, called

iLoc-Gpos, has been developed for predicting the subcellular localization of Gram

positive bacterial proteins with both single-location and multiple-location

sites. For facilitating comparison, the same stringent benchmark dataset used to

estimate the accuracy of Gpos-mPLoc was adopted to demonstrate the power of

iLoc-Gpos. The dataset contains 519 Gram-positive bacterial proteins classified

into the following four subcellular locations: (1) cell membrane, (2) cell wall,

(3) cytoplasm, and (4) extracell; none of proteins included has ≥25% pairwise

sequence identity to any other in a same subset (subcellular location). The

overall success rate by jackknife test on such a stringent benchmark dataset by

iLoc-Gpos was over 93%, which is about 11% higher than that by GposmPLoc. As a

user-friendly web-server, iLoc-Gpos is freely accessible to the public at

http://icpr.jci.edu.cn/bioinfo/iLoc- Gpos or http://www.jci-bioinfo.cn/iLoc-Gpos.

Meanwhile, a step-by-step guide is provided on how to use the web-server to get

the desired results. Furthermore, for the user � s convenience, the iLoc-Gpos

web-server also has the function to accept the batch job submission, which is not

available in the existing version of Gpos-mPLoc web-server.

PMID: 21919865 [Indexed for MEDLINE]

1550. Protein Pept Lett. 2012 Jan;19(1):15-22.

PSCL: predicting protein subcellular localization based on optimal functional

domains.

Wang K(1), Hu LL, Shi XH, Dong YS, Li HP, Wen TQ.

Author information:

(1)Laboratory of Molecular Neurobiology, School of Life Sciences, Shanghai

University, Shanghai, PR China.

It is well known that protein subcellular localizations are closely related to

their functions. Although many computational methods and tools are available from

Internet, it is still necessary to develop new algorithms in this filed to gain a

better understanding of the complex mechanism of plant subcellular localization.

Here, we provide a new web server named PSCL for plant protein subcellular

localization prediction by employing optimized functional domains. After feature

optimization, 848 optimal functional domains from InterPro were obtained to

represent each protein. By calculating the distances to each of the seven

categories, PSCL showing the possibilities of a protein located into each of

those categories in ascending order. Toward our dataset, PSCL achieved a

first-order predicted accuracy of 75.7% by jackknife test. Gene Ontology

enrichment analysis showing that catalytic activity, cellular process and

metabolic process are strongly correlated with the localization of plant

proteins. Finally, PSCL, a Linux Operate System based web interface for the

predictor was designed and is accessible for public use at

http://pscl.biosino.org/.

PMID: 21919864 [Indexed for MEDLINE]

1551. Protein Pept Lett. 2012 Jan;19(1):57-61.

SCYPPred: a web-based predictor of SNPs for human cytochrome P450.

Li L(1), Wei DQ, Wang JF, Chou KC.

Author information:

(1)State Key Laboratory of Microbial Metabolism, College of Life Sciences and

Biotechnology, Shanghai Jiao Tong University, Shanghai, P R China.

Human cytochrome P450(CYP 450) enzymes mediate over 60% of the phase I-dependent

metabolism of clinical drugs. They are also known for the polymorphism functions

that have significant impacts on the enzyme activities. In this study, a

web-server called SCYPPred was developed for predicting human cytochrome P450

SNPs (Single Nucleotide Polymorphisms) based on the SVM flanking sequence method;

SCYPPred can rapidly yield the desired results by using the amino acid sequences

information alone. The web-server is accessible to the public at

http://snppred.sjtu.edu.cn. Hopefully SCYPPred could be a useful bioinformatics

tool for elucidating the mutation probability of a specific CYP450 enzyme.

PMID: 21919859 [Indexed for MEDLINE]

1552. Proteins. 2012 Jan;80(1):93-110. doi: 10.1002/prot.23165. Epub 2011 Oct 4.

BSP-SLIM: a blind low-resolution ligand-protein docking approach using predicted

protein structures.

Lee HS(1), Zhang Y.

Author information:

(1)Department of Biological Chemistry, Center for Computational Medicine and

Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA.

We developed BSP-SLIM, a new method for ligand-protein blind docking using

low-resolution protein structures. For a given sequence, protein structures are

first predicted by I-TASSER; putative ligand binding sites are transferred from

holo-template structures which are analogous to the I-TASSER models;

ligand-protein docking conformations are then constructed by shape and chemical

match of ligand with the negative image of binding pockets. BSP-SLIM was tested

on 71 ligand-protein complexes from the Astex diverse set where the protein

structures were predicted by I-TASSER with an average RMSD 2.92 Å on the binding

residues. Using I-TASSER models, the median ligand RMSD of BSP-SLIM docking is

3.99 Å which is 5.94 Å lower than that by AutoDock; the median binding-site error

by BSP-SLIM is 1.77 Å which is 6.23 Å lower than that by AutoDock and 3.43 Å

lower than that by LIGSITE(CSC) . Compared to the models using crystal protein

structures, the median ligand RMSD by BSP-SLIM using I-TASSER models increases by

0.87 Å, while that by AutoDock increases by 8.41 Å; the median binding-site error

by BSP-SLIM increase by 0.69Å while that by AutoDock and LIGSITE(CSC) increases

by 7.31 Å and 1.41 Å, respectively. As case studies, BSP-SLIM was used in virtual

screening for six target proteins, which prioritized actives of 25% and 50% in

the top 9.2% and 17% of the library on average, respectively. These results

demonstrate the usefulness of the template-based coarse-grained algorithms in the

low-resolution ligand-protein docking and drug-screening. An on-line BSP-SLIM

server is freely available at http://zhanglab.ccmb.med.umich.edu/BSP-SLIM.

Copyright © 2011 Wiley Periodicals, Inc.

DOI: 10.1002/prot.23165

PMCID: PMC3240723

PMID: 21971880 [Indexed for MEDLINE]

1553. Stud Health Technol Inform. 2012;180:270-4.

Automated realtime data import for the i2b2 clinical data warehouse: introducing

the HL7 ETL cell.

Majeed RW(1), Röhrig R.

Author information:

(1)Section for Medical Informatics in Anesthesiology, Justus Liebig University

Gießen, Germany.

Clinical data warehouses are used to consolidate all available clinical data from

one or multiple organizations. They represent an important source for clinical

research, quality management and controlling. Since its introduction, the data

warehouse i2b2 gathered a large user base in the research community. Yet, little

work has been done on the process of importing clinical data into data warehouses

using existing standards. In this article, we present a novel approach of

utilizing the clinical integration server as data source, commonly available in

most hospitals. As information is transmitted through the integration server, the

standardized HL7 message is immediately parsed and inserted into the data

warehouse. Evaluation of import speeds suggest feasibility of the provided

solution for real-time processing of HL7 messages. By using the presented

approach of standardized data import, i2b2 can be used as a plug and play data

warehouse, without the hurdle of customized import for every clinical information

system or electronic medical record. The provided solution is available for

download at http://sourceforge.net/projects/histream/.

PMID: 22874194 [Indexed for MEDLINE]

1554. J Chem Inf Model. 2011 Dec 27;51(12):3078-92. doi: 10.1021/ci200377u. Epub 2011

Nov 21.

Statistical potential for modeling and ranking of protein-ligand interactions.

Fan H(1), Schneidman-Duhovny D, Irwin JJ, Dong G, Shoichet BK, Sali A.

Author information:

(1)Department of Bioengineering and Therapeutic Sciences, University of

California, San Francisco, USA.

Applications in structural biology and medicinal chemistry require protein-ligand

scoring functions for two distinct tasks: (i) ranking different poses of a small

molecule in a protein binding site and (ii) ranking different small molecules by

their complementarity to a protein site. Using probability theory, we developed

two atomic distance-dependent statistical scoring functions: PoseScore was

optimized for recognizing native binding geometries of ligands from other poses

and RankScore was optimized for distinguishing ligands from nonbinding molecules.

Both scores are based on a set of 8,885 crystallographic structures of

protein-ligand complexes but differ in the values of three key parameters.

Factors influencing the accuracy of scoring were investigated, including the

maximal atomic distance and non-native ligand geometries used for scoring, as

well as the use of protein models instead of crystallographic structures for

training and testing the scoring function. For the test set of 19 targets,

RankScore improved the ligand enrichment (logAUC) and early enrichment (EF(1))

scores computed by DOCK 3.6 for 13 and 14 targets, respectively. In addition,

RankScore performed better at rescoring than each of seven other scoring

functions tested. Accepting both the crystal structure and decoy geometries with

all-atom root-mean-square errors of up to 2 Å from the crystal structure as

correct binding poses, PoseScore gave the best score to a correct binding pose

among 100 decoys for 88% of all cases in a benchmark set containing 100

protein-ligand complexes. PoseScore accuracy is comparable to that of

DrugScore(CSD) and ITScore/SE and superior to 12 other tested scoring functions.

Therefore, RankScore can facilitate ligand discovery, by ranking complexes of the

target with different small molecules; PoseScore can be used for protein-ligand

complex structure prediction, by ranking different conformations of a given

protein-ligand pair. The statistical potentials are available through the

Integrative Modeling Platform (IMP) software package (http://salilab.org/imp) and

the LigScore Web server (http://salilab.org/ligscore/).

DOI: 10.1021/ci200377u

PMCID: PMC3246566

PMID: 22014038 [Indexed for MEDLINE]

1555. BMC Bioinformatics. 2011 Dec 22;12:489. doi: 10.1186/1471-2105-12-489.

Predicting RNA-protein interactions using only sequence information.

Muppirala UK(1), Honavar VG, Dobbs D.

Author information:

(1)Bioinformatics and Computational Biology Program, Iowa State University, Ames,

Iowa, USA. usha@iastate.edu

BACKGROUND: RNA-protein interactions (RPIs) play important roles in a wide

variety of cellular processes, ranging from transcriptional and

post-transcriptional regulation of gene expression to host defense against

pathogens. High throughput experiments to identify RNA-protein interactions are

beginning to provide valuable information about the complexity of RNA-protein

interaction networks, but are expensive and time consuming. Hence, there is a

need for reliable computational methods for predicting RNA-protein interactions.

RESULTS: We propose RPISeq, a family of classifiers for predicting RNA-protein

interactions using only sequence information. Given the sequences of an RNA and a

protein as input, RPIseq predicts whether or not the RNA-protein pair interact.

The RNA sequence is encoded as a normalized vector of its ribonucleotide 4-mer

composition, and the protein sequence is encoded as a normalized vector of its

3-mer composition, based on a 7-letter reduced alphabet representation. Two

variants of RPISeq are presented: RPISeq-SVM, which uses a Support Vector Machine

(SVM) classifier and RPISeq-RF, which uses a Random Forest classifier. On two

non-redundant benchmark datasets extracted from the Protein-RNA Interface

Database (PRIDB), RPISeq achieved an AUC (Area Under the Receiver Operating

Characteristic (ROC) curve) of 0.96 and 0.92. On a third dataset containing only

mRNA-protein interactions, the performance of RPISeq was competitive with that of

a published method that requires information regarding many different features

(e.g., mRNA half-life, GO annotations) of the putative RNA and protein partners.

In addition, RPISeq classifiers trained using the PRIDB data correctly predicted

the majority (57-99%) of non-coding RNA-protein interactions in NPInter-derived

networks from E. coli, S. cerevisiae, D. melanogaster, M. musculus, and H.

sapiens.

CONCLUSIONS: Our experiments with RPISeq demonstrate that RNA-protein

interactions can be reliably predicted using only sequence-derived information.

RPISeq offers an inexpensive method for computational construction of RNA-protein

interaction networks, and should provide useful insights into the function of

non-coding RNAs. RPISeq is freely available as a web-based server at

http://pridb.gdcb.iastate.edu/RPISeq/.

DOI: 10.1186/1471-2105-12-489

PMCID: PMC3322362

PMID: 22192482 [Indexed for MEDLINE]

1556. Bioinformatics. 2011 Dec 15;27(24):3432-4. doi: 10.1093/bioinformatics/btr582.

Epub 2011 Oct 20.

VizPrimer: a web server for visualized PCR primer design based on known gene

structure.

Zhou Y(1), Qu W, Lu Y, Zhang Y, Wang X, Zhao D, Yang Y, Zhang C.

Author information:

(1)Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics,

Cognitive and Mental Health Research Center, Beijing 100850, China.

SUMMARY: The visualization of gene structure plays an important role in

polymerase chain reaction (PCR) primer design, especially for eukaryotic genes

with a number of splice variants that users need to distinguish between via PCR.

Here, we describe a visualized web server for primer design named VizPrimer. It

utilizes the new information technology (IT) tools, HTML5 to display gene

structure and JavaScript to interact with the users. In VizPrimer, the users can

focus their attention on the gene structure and primer design strategy, without

wasting time calculating the exon positions of splice variants or manually

configuring complicated parameters. In addition, VizPrimer is also suitable for

the design of PCR primers for amplifying open reading frames and detecting single

nucleotide polymorphisms (SNPs).

AVAILABILITY: VizPrimer is freely available at

http://biocompute.bmi.ac.cn/CZlab/VizPrimer/. The web server supported browsers:

Chrome (≥5.0), Firefox (≥3.0), Safari (≥4.0) and Opera (≥10.0).

CONTACT: zhangcg@bmi.ac.cn; yangyi528@vip.sina.com.

DOI: 10.1093/bioinformatics/btr582

PMID: 22016409 [Indexed for MEDLINE]

1557. Bioinformatics. 2011 Dec 15;27(24):3333-40. doi: 10.1093/bioinformatics/btr570.

Epub 2011 Oct 12.

Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data.

Althammer S(1), González-Vallinas J, Ballaré C, Beato M, Eyras E.

Author information:

(1)Universitat Pompeu Fabra, E08003 Barcelona, Spain.

MOTIVATION: High-throughput sequencing (HTS) has revolutionized gene regulation

studies and is now fundamental for the detection of protein-DNA and protein-RNA

binding, as well as for measuring RNA expression. With increasing variety and

sequencing depth of HTS datasets, the need for more flexible and memory-efficient

tools to analyse them is growing.

RESULTS: We describe Pyicos, a powerful toolkit for the analysis of mapped reads

from diverse HTS experiments: ChIP-Seq, either punctuated or broad signals,

CLIP-Seq and RNA-Seq. We prove the effectiveness of Pyicos to select for

significant signals and show that its accuracy is comparable and sometimes

superior to that of methods specifically designed for each particular type of

experiment. Pyicos facilitates the analysis of a variety of HTS datatypes through

its flexibility and memory efficiency, providing a useful framework for data

integration into models of regulatory genomics.

AVAILABILITY: Open-source software, with tutorials and protocol files, is

available at http://regulatorygenomics.upf.edu/pyicos or as a Galaxy server at

http://regulatorygenomics.upf.edu/galaxy

CONTACT: eduardo.eyras@upf.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr570

PMCID: PMC3232367

PMID: 21994224 [Indexed for MEDLINE]

1558. BMC Bioinformatics. 2011 Dec 14;12 Suppl 14:S6. doi: 10.1186/1471-2105-12-S14-S6.

iGepros: an integrated gene and protein annotation server for biological nature

exploration.

Zheng G(1), Wang H, Wei C, Li Y.

Author information:

(1)Key Laboratory of Systems Biology, Shanghai Institutes for Biological

Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China.

zhenggy@sibs.ac.cn

BACKGROUND: In the post-genomic era, transcriptomics and proteomics provide

important information to understand the genomes. With fast development of

high-throughput technology, more and more transcriptomics and proteomics data are

generated at an unprecedented rate. Therefore, requirement of software to

annotate those omics data and explore their biological nature arises. In the past

decade, some pioneer works were presented to address this issue, but limitations

still exist. Fox example, some of these tools offer command line only, which is

not suitable for those users with little or no experience in programming.

Besides, some tools don't support large scale gene and protein analysis.

RESULTS: To overcome these limitations, an integrated gene and protein annotation

server named iGepros has been developed. The server provides user-friendly

interfaces and detailed on-line examples, so most researchers even those with

little or no programming experience can use it smoothly. Moreover, the server

provides many functionalities to compare transcriptomics and proteomics data.

Especially, the server is constructed under a model-view-control framework, which

makes it easy to incorporate more functions to the server in the future.

CONCLUSIONS: In this paper, we present a server with powerful capability not only

for gene and protein functional annotation, but also for transcriptomics and

proteomics data comparison. Researchers can survey biological characters behind

gene and protein datasets and accelerate their investigation of transcriptome and

proteome by applying the server. The server is publicly available at

http://www.biosino.org/iGepros/.

DOI: 10.1186/1471-2105-12-S14-S6

PMCID: PMC3287471

PMID: 22373022 [Indexed for MEDLINE]

1559. J Chem Theory Comput. 2011 Dec 13;7(12):4026-37. doi: 10.1021/ct200196m. Epub

2011 Nov 15.

An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0.

Malde AK(1), Zuo L(1), Breeze M(1), Stroet M(1), Poger D(1), Nair PC(1),

Oostenbrink C(2,)(3), Mark AE(1,)(4).

Author information:

(1)School of Chemistry and Molecular Biosciences, University of Queensland , St.

Lucia, Australia. (2)Leiden/Amsterdam Center for Drug Research, Division of

Molecular Toxicology, Vrije Universiteit Amsterdam , Amsterdam, The Netherlands.

(3)Institute of Molecular Modeling and Simulation, University of Natural

Resources and Life Sciences , Vienna, Austria. (4)Institute for Molecular

Bioscience, University of Queensland , St. Lucia, QLD 4072, Australia.

The Automated force field Topology Builder (ATB,

http://compbio.biosci.uq.edu.au/atb ) is a Web-accessible server that can provide

topologies and parameters for a wide range of molecules appropriate for use in

molecular simulations, computational drug design, and X-ray refinement. The ATB

has three primary functions: (1) to act as a repository for molecules that have

been parametrized as part of the GROMOS family of force fields, (2) to act as a

repository for pre-equilibrated systems for use as starting configurations in

molecular dynamics simulations (solvent mixtures, lipid systems pre-equilibrated

to adopt a specific phase, etc.), and (3) to generate force field descriptions of

novel molecules compatible with the GROMOS family of force fields in a variety of

formats (GROMOS, GROMACS, and CNS). Force field descriptions of novel molecules

are derived using a multistep process in which results from quantum mechanical

(QM) calculations are combined with a knowledge-based approach to ensure

compatibility (as far as possible) with a specific parameter set of the GROMOS

force field. The ATB has several unique features: (1) It requires that the user

stipulate the protonation and tautomeric states of the molecule. (2) The symmetry

of the molecule is analyzed to ensure that equivalent atoms are assigned

identical parameters. (3) Charge groups are assigned automatically. (4) Where the

assignment of a given parameter is ambiguous, a range of possible alternatives is

provided. The ATB also provides several validation tools to assist the user to

assess the degree to which the topology generated may be appropriate for a given

task. In addition to detailing the steps involved in generating a force field

topology compatible with a specific GROMOS parameter set (GROMOS 53A6), the

challenges involved in the automatic generation of force field parameters for

atomic simulations in general are discussed.

DOI: 10.1021/ct200196m

PMID: 26598349

1560. Structure. 2011 Dec 7;19(12):1744-51. doi: 10.1016/j.str.2011.10.015.

Automated prediction of protein association rate constants.

Qin S(1), Pang X, Zhou HX.

Author information:

(1)Department of Physics and Institute of Molecular Biophysics, Tallahassee, FL

32306, USA.

The association rate constants (k(a)) of proteins with other proteins or other

macromolecular targets are a fundamental biophysical property. Observed rate

constants span over ten orders of magnitude, from 1 to 10(10) M(-1)s(-1). Protein

association can be rate limited either by the diffusional approach of the

subunits to form a transient complex, with near-native separation and orientation

but without short-range native interactions, or by the subsequent conformational

rearrangement to form the native complex. Our transient-complex theory showed

promise in predicting k(a) in the diffusion-limited regime. Here, we develop it

into a web server called TransComp (http://pipe.sc.fsu.edu/transcomp/) and report

on the server's accuracy and robustness based on applications to over 100 protein

complexes. We expect this server to be a valuable tool for systems biology

applications and for kinetic characterization of protein-protein and

protein-nucleic acid association in general.

Copyright © 2011 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.str.2011.10.015

PMCID: PMC3240845

PMID: 22153497 [Indexed for MEDLINE]

1561. Bioinformatics. 2011 Dec 1;27(23):3286-92. doi: 10.1093/bioinformatics/btr576.

Epub 2011 Oct 13.

Protein stability: a single recorded mutation aids in predicting the effects of

other mutations in the same amino acid site.

Wainreb G(1), Wolf L, Ashkenazy H, Dehouck Y, Ben-Tal N.

Author information:

(1)Department of Biochemistry and Molecular Biology, Tel-Aviv University, Ramat

Aviv 69978, Israel.

MOTIVATION: Accurate prediction of protein stability is important for

understanding the molecular underpinnings of diseases and for the design of new

proteins. We introduce a novel approach for the prediction of changes in protein

stability that arise from a single-site amino acid substitution; the approach

uses available data on mutations occurring in the same position and in other

positions. Our algorithm, named Pro-Maya (Protein Mutant stAbilitY Analyzer),

combines a collaborative filtering baseline model, Random Forests regression and

a diverse set of features. Pro-Maya predicts the stability free energy difference

of mutant versus wild type, denoted as ΔΔG.

RESULTS: We evaluated our algorithm extensively using cross-validation on two

previously utilized datasets of single amino acid mutations and a (third)

validation set. The results indicate that using known ΔΔG values of mutations at

the query position improves the accuracy of ΔΔG predictions for other mutations

in that position. The accuracy of our predictions in such cases significantly

surpasses that of similar methods, achieving, e.g. a Pearson's correlation

coefficient of 0.79 and a root mean square error of 0.96 on the validation set.

Because Pro-Maya uses a diverse set of features, including predictions using two

other methods, it also performs slightly better than other methods in the absence

of additional experimental data on the query positions.

AVAILABILITY: Pro-Maya is freely available via web server at

http://bental.tau.ac.il/ProMaya.

CONTACT: nirb@tauex.tau.ac.il; wolf@cs.tau.ac.il

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr576

PMCID: PMC3223369

PMID: 21998155 [Indexed for MEDLINE]

1562. Bioinformatics. 2011 Dec 1;27(23):3315-6. doi: 10.1093/bioinformatics/btr575.

Epub 2011 Oct 12.

Automatic generation of protein structure cartoons with Pro-origami.

Stivala A(1), Wybrow M, Wirth A, Whisstock JC, Stuckey PJ.

Author information:

(1)Department of Computer Science and Software Engineering, The University of

Melbourne Parkville Campus, Victoria 3010, Australia.

a.stivala@pgrad.unimelb.edu.au

SUMMARY: Protein topology diagrams are 2D representations of protein structure

that are particularly useful in understanding and analysing complex protein

folds. Generating such diagrams presents a major problem in graph drawing, with

automatic approaches often resulting in errors or uninterpretable results. Here

we apply a breakthrough in diagram layout to protein topology cartoons, providing

clear, accurate, interactive and editable diagrams, which are also an interface

to a structural search method.

AVAILABILITY: Pro-origami is available via a web server at

http://munk.csse.unimelb.edu.au/pro-origami

CONTACT: a.stivala@pgrad.unimelb.edu.au; pjs@csse.unimelb.edu.au.

DOI: 10.1093/bioinformatics/btr575

PMID: 21994221 [Indexed for MEDLINE]

1563. Bioinformatics. 2011 Dec 1;27(23):3321-2. doi: 10.1093/bioinformatics/btr557.

Epub 2011 Oct 7.

MyBioNet: interactively visualize, edit and merge biological networks on the web.

Huang D(1), Huang Y, Bai Y, Chen D, Hofestädt R, Klukas C, Chen M.

Author information:

(1)Department of Bioinformatics, State Key Laboratory of Plant Physiology and

Biochemistry, College of Life Sciences, Zhejiang University, Hangzhou 310058,

China.

SUMMARY: MyBioNet is a web-based application for biological network analysis,

which provides user-friendly web interfaces to visualize, edit and merge

biological networks. In addition, MyBioNet integrated KEGG metabolic network data

from 1366 organisms and allows users to search and navigate interesting networks.

AVAILABILITY AND IMPLEMENTATION: All KEGG metabolic network data are organized

and stored in the MySQL database. MyBioNet is implemented in Flex/Actionscript

and PHP languages and deployed on an Apache web server. MyBioNet is accessible

through all the Flash-embedded browsers at http://bis.zju.edu.cn/mybionet/.

CONTACT: mchen@zju.edu.cn.

DOI: 10.1093/bioinformatics/btr557

PMID: 21984760 [Indexed for MEDLINE]

1564. Bioinformatics. 2011 Dec 1;27(23):3317-8. doi: 10.1093/bioinformatics/btr548.

Epub 2011 Oct 3.

ProfileChaser: searching microarray repositories based on genome-wide patterns of

differential expression.

Engreitz JM(1), Chen R, Morgan AA, Dudley JT, Mallelwar R, Butte AJ.

Author information:

(1)Division of Systems Medicine, Department of Pediatrics, Stanford University

School of Medicine, Stanford, CA 94305, USA.

SUMMARY: We introduce ProfileChaser, a web server that allows for querying the

Gene Expression Omnibus based on genome-wide patterns of differential expression.

Using a novel, content-based approach, ProfileChaser retrieves expression

profiles that match the differentially regulated transcriptional programs in a

user-supplied experiment. This analysis identifies statistical links to similar

expression experiments from the vast array of publicly available data on

diseases, drugs, phenotypes and other experimental conditions.

AVAILABILITY: http://profilechaser.stanford.edu

CONTACT: abutte@stanford.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr548

PMCID: PMC3223361

PMID: 21967760 [Indexed for MEDLINE]

1565. Curr Protoc Bioinformatics. 2011 Dec;Chapter 13:Unit 13.5.. doi:

10.1002/0471250953.bi1305s36.

Installation and use of LabKey Server for proteomics.

Eckels J(1), Hussey P, Nelson EK, Myers T, Rauch A, Bellew M, Connolly B, Law W,

Eng JK, Katz J, McIntosh M, Mallick P, Igra M.

Author information:

(1)LabKey Software, Seattle, Washington, USA.

LabKey Server (formerly CPAS, the Computational Proteomics Analysis System)

provides a Web-based platform for mining data from liquid chromatography-tandem

mass spectrometry (LC-MS/MS) proteomic experiments. This open source platform

supports systematic proteomic analyses and secure data management, integration,

and sharing. LabKey Server incorporates several tools currently used in proteomic

analysis, including the X! Tandem search engine, the ProteoWizard toolkit, and

the PeptideProphet and ProteinProphet data mining tools. These tools and others

are integrated into LabKey Server, which provides an extensible architecture for

developing high-throughput biological applications. The LabKey Server analysis

pipeline acts on data in standardized file formats, so that researchers may use

LabKey Server with other search engines, including Mascot or SEQUEST, that follow

a standardized format for reporting search engine results. Supported builds of

LabKey Server are freely available at http://www.labkey.com/. Documentation and

source code are available under the Apache License 2.0 at http://www.labkey.org.

© 2011 by John Wiley & Sons, Inc.

DOI: 10.1002/0471250953.bi1305s36

PMID: 22161569 [Indexed for MEDLINE]

1566. DNA Res. 2011 Dec;18(6):463-70. doi: 10.1093/dnares/dsr032. Epub 2011 Oct 10.

PlantRGS: a web server for the identification of most suitable candidate

reference genes for quantitative gene expression studies in plants.

Patel RK(1), Jain M.

Author information:

(1)National Institute of Plant Genome Research (NIPGR), Aruna Asaf Ali Marg, New

Delhi, India.

Normalization of quantitative gene expression data with a suitable reference gene

is essential for accurate and reliable results. However, the availability and

choice of most suitable reference gene(s) showing uniform expression across all

the experimental conditions remain a drawback. We have developed a web server,

PlantRGS (http://www.nipgr.res.in/PlantRGS), for the identification of most

suitable candidate reference gene(s) at the whole-genome level using microarray

data for quantitative gene expression studies in plants. Microarray data from

more than 11 000 tissue samples for nine plant species have been included in the

PlantRGS for meta-analysis. The web server provides a user-friendly graphical

user interface-based analysis tool for the identification of most suitable

reference genes in the selected plant species under user-defined experimental

conditions. Various parameter options and output formats will help users to

investigate desired number of most suitable reference genes with wide range of

expression levels. Validation of results revealed that novel reference genes

identified by the PlantRGS outperforms the traditionally used reference genes in

terms of expression stability. We anticipate that the PlantRGS will provide a

platform for the identification of most suitable reference gene(s) under given

experimental conditions and facilitate quantitative gene expression studies in

plants.

DOI: 10.1093/dnares/dsr032

PMCID: PMC3223078

PMID: 21987088 [Indexed for MEDLINE]

1567. Epigenetics. 2011 Dec;6(12):1505-12. doi: 10.4161/epi.6.12.18176.

EpiRegNet: constructing epigenetic regulatory network from high throughput gene

expression data for humans.

Wang LY(1), Wang P, Li MJ, Qin J, Wang X, Zhang MQ, Wang J.

Author information:

(1)Department of Biochemistry, LKS Faculty of Medicine, The University of Hong

Kong, Hong Kong, China.

The advances of high throughput profiling methods, such as microarray gene

profiling and RNA-seq, have enabled researchers to identify thousands of

differentially expressed genes under a certain perturbation. Much work has been

done to understand the genetic factors that contribute to the expression changes

by searching the over-represented regulatory motifs in the promoter regions of

these genes. However, the changes could also be caused by epigenetic regulation,

especially histone modifications, and no web server has been constructed to study

the epigenetic factors responsible for gene expression changes. Here, we pre-sent

a web tool for this purpose. Provided with different categories of genes (e.g.,

up-regulated, down-regulated or unchanged genes), the server will find epigenetic

factors responsible for the difference among the categories and construct an

epigenetic regulatory network. Furthermore, it will perform co-localization

analyses between these epigenetic factors and transcription factors, which were

collected from large scale experimental ChIP-seq or computational predicted data.

In addition, for users who want to analyze dynamic change of a histone

modification mark under different cell conditions, the server will find direct

and indirect target genes of this mark by integrative analysis of experimental

data and computational prediction, and present a regulatory network around this

mark. Both networks can be visualized by a user friendly interface and the data

are downloadable in batch. The server currently supports 12 cell types in human,

including ESC and CD4+ T cells, and will expand as more public data are

available. It also allows user to create a self-defined cell type, upload and

analyze multiple ChIP-seq data. It is freely available to academic users at

http://jjwanglab.org/EpiRegNet.

DOI: 10.4161/epi.6.12.18176

PMID: 22139581 [Indexed for MEDLINE]

1568. IEEE Trans Nanobioscience. 2011 Dec;10(4):248-9. doi: 10.1109/TNB.2011.2169331.

Epub 2011 Sep 23.

3-d brownian motion simulator for high-sensitivity nanobiotechnological

applications.

Toth A(1), Banky D, Grolmusz V.

Author information:

(1)Mathematical Institute of Eötvös University, Budapest, Hungary.

A wide variety of nanobiotechnologic applications are being developed for

nanoparticle based in vitro diagnostic and imaging systems. Some of these systems

make possible highly sensitive detection of molecular biomarkers. Frequently, the

very low concentration of the biomarkers makes impossible the classical, partial

differential equation-based mathematical simulation of the motion of the

nanoparticles involved. We present a three-dimensional Brownian motion simulation

tool for the prediction of the movement of nanoparticles in various thermal,

viscosity, and geometric settings in a rectangular cuvette. For nonprofit users

the server is freely available at the site http://brownian.pitgroup.org.

DOI: 10.1109/TNB.2011.2169331

PMID: 21947531 [Indexed for MEDLINE]

1569. Infect Genet Evol. 2011 Dec;11(8):2151-61. doi: 10.1016/j.meegid.2011.09.021.

Epub 2011 Oct 1.

Identification of Salmonella enterica species- and subgroup-specific genomic

regions using Panseq 2.0.

Laing C(1), Villegas A, Taboada EN, Kropinski A, Thomas JE, Gannon VP.

Author information:

(1)Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, Lethbridge,

AB, Canada. chad.r.laing@phac-aspc.gc.ca

The pan-genome of a taxonomic group consists of evolutionarily conserved core

genes shared by all members and accessory genes that are present only in some

members of the group. Group- and subgroup-specific core genes are thought to

contribute to shared phenotypes such as virulence and niche specificity. In this

study we analyzed 39 Salmonella enterica genomes (16 closed, 23 draft), a species

that contains two human-specific serovars that cause typhoid fever, as well as a

large number of zoonotic serovars that cause gastroenteritis in humans. Panseq

2.0 was used to define the pan-genome by adjusting the threshold at which

group-specific "core" loci are defined. We found the pan-genome to be 9.03 Mbp in

size, and that the core genome size decreased, while the number of SNPs/100 bp

increased, as the number of strains used to define the core genome increased,

suggesting substantial divergence among S. enterica subgroups. Subgroup-specific

"core" genes, in contrast, had fewer SNPs/100 bp, likely reflecting their more

recent acquisition. Phylogenetic trees were created from the concatenated and

aligned pan-genome, the core genome, and multi-locus-sequence typing (MLST) loci.

Branch support increased among the trees, and strains of the same serovar grouped

closer together as the number of loci used to create the tree increased. Further,

high levels of discrimination were achieved even amongst the most closely related

strains of S. enterica Typhi, suggesting that the data generated by Panseq may

also be of value in short-term epidemiological studies. Panseq provides an easy

and fast way of performing pan-genomic analyses, which can include the

identification of group-dominant as well as group-specific loci and is available

as a web-server and a standalone version at http://lfz.corefacility.ca/panseq/.

Crown Copyright © 2011. Published by Elsevier B.V. All rights reserved.

DOI: 10.1016/j.meegid.2011.09.021

PMID: 22001825 [Indexed for MEDLINE]

1570. J Struct Funct Genomics. 2011 Dec;12(4):181-9. doi: 10.1007/s10969-011-9119-x.

Epub 2011 Dec 3.

PSS-3D1D: an improved 3D1D profile method of protein fold recognition for the

annotation of twilight zone sequences.

Ganesan K(1), Parthasarathy S.

Author information:

(1)Department of Bioinformatics, Bharathidasan University, Tiruchirappalli, Tamil

Nadu, India.

Annotation of any newly determined protein sequence depends on the pairwise

sequence identity with known sequences. However, for the twilight zone sequences

which have only 15-25% identity, the pair-wise comparison methods are inadequate

and the annotation becomes a challenging task. Such sequences can be annotated by

using methods that recognize their fold. Bowie et al. described a 3D1D profile

method in which the amino acid sequences that fold into a known 3D structure are

identified by their compatibility to that known 3D structure. We have improved

the above method by using the predicted secondary structure information and

employ it for fold recognition from the twilight zone sequences. In our Protein

Secondary Structure 3D1D (PSS-3D1D) method, a score (w) for the predicted

secondary structure of the query sequence is included in finding the

compatibility of the query sequence to the known fold 3D structures. In the

benchmarks, the PSS-3D1D method shows a maximum of 21% improvement in predicting

correctly the α + β class of folds from the sequences with twilight zone level of

identity, when compared with the 3D1D profile method. Hence, the PSS-3D1D method

could offer more clues than the 3D1D method for the annotation of twilight zone

sequences. The web based PSS-3D1D method is freely available in the PredictFold

server at http://bioinfo.bdu.ac.in/servers/ .

DOI: 10.1007/s10969-011-9119-x

PMID: 22160493 [Indexed for MEDLINE]

1571. Nucleic Acids Res. 2011 Dec;39(22):e153. doi: 10.1093/nar/gkr770. Epub 2011 Oct

5.

Prioritizing human cancer microRNAs based on genes' functional consistency

between microRNA and cancer.

Li X(1), Wang Q, Zheng Y, Lv S, Ning S, Sun J, Huang T, Zheng Q, Ren H, Xu J,

Wang X, Li Y.

Author information:

(1)College of Bioinformatics Science and Technology, Cancer Hospital Affiliated

to Harbin Medical University, Harbin, 150081, China. lixia@hrbmu.edu.cn

The identification of human cancer-related microRNAs (miRNAs) is important for

cancer biology research. Although several identification methods have achieved

remarkable success, they have overlooked the functional information associated

with miRNAs. We present a computational framework that can be used to prioritize

human cancer miRNAs by measuring the association between cancer and miRNAs based

on the functional consistency score (FCS) of the miRNA target genes and the

cancer-related genes. This approach proved successful in identifying the

validated cancer miRNAs for 11 common human cancers with area under ROC curve

(AUC) ranging from 71.15% to 96.36%. The FCS method had a significant advantage

over miRNA differential expression analysis when identifying cancer-related

miRNAs with a fine regulatory mechanism, such as miR-27a in colorectal cancer.

Furthermore, a case study examining thyroid cancer showed that the FCS method can

uncover novel cancer-related miRNAs such as miR-27a/b, which were showed

significantly upregulated in thyroid cancer samples by qRT-PCR analysis. Our

method can be used on a web-based server, CMP (cancer miRNA prioritization) and

is freely accessible at http://bioinfo.hrbmu.edu.cn/CMP. This time- and

cost-effective computational framework can be a valuable complement to

experimental studies and can assist with future studies of miRNA involvement in

the pathogenesis of cancers.

© The Author(s) 2011. Published by Oxford University Press.

DOI: 10.1093/nar/gkr770

PMCID: PMC3239203

PMID: 21976726 [Indexed for MEDLINE]

1572. BMC Genomics. 2011 Nov 30;12 Suppl 3:S9. doi: 10.1186/1471-2164-12-S3-S9. Epub

2011 Nov 30.

UASIS: Universal Automatic SNP Identification System.

Poo DC(1), Cai S, Mah JT.

Author information:

(1)Department of Information Systems, School of Computing, National University of

Singapore, 13 Computing Drive, Singapore 117417. dannypoo@nus.edu.sg

BACKGROUND: SNP (Single Nucleotide Polymorphism), the most common genetic

variations between human beings, is believed to be a promising way towards

personalized medicine. As more and more research on SNPs are being conducted,

non-standard nomenclatures may generate potential problems. The most serious

issue is that researchers cannot perform cross referencing among different SNP

databases. This will result in more resources and time required to track SNPs. It

could be detrimental to the entire academic community.

RESULTS: UASIS (Universal Automated SNP Identification System) is a web-based

server for SNP nomenclature standardization and translation at DNA level. Three

utilities are available. They are UASIS Aligner, Universal SNP Name Generator and

SNP Name Mapper. UASIS maps SNPs from different databases, including dbSNP, GWAS,

HapMap and JSNP etc., into an uniform view efficiently using a proposed universal

nomenclature and state-of-art alignment algorithms. UASIS is freely available at

http://www.uasis.tk with no requirement of log-in.

CONCLUSIONS: UASIS is a helpful platform for SNP cross referencing and tracking.

By providing an informative, unique and unambiguous nomenclature, which utilizes

unique position of a SNP, we aim to resolve the ambiguity of SNP nomenclatures

currently practised. Our universal nomenclature is a good complement to

mainstream SNP notations such as rs# and HGVS guidelines. UASIS acts as a bridge

to connect heterogeneous representations of SNPs.

DOI: 10.1186/1471-2164-12-S3-S9

PMCID: PMC3333510

PMID: 22369494 [Indexed for MEDLINE]

1573. BMC Genomics. 2011 Nov 30;12 Suppl 3:S12. doi: 10.1186/1471-2164-12-S3-S12. Epub

2011 Nov 30.

i-rDNA: alignment-free algorithm for rapid in silico detection of ribosomal gene

fragments from metagenomic sequence data sets.

Mohammed MH(1), Ghosh TS, Chadaram S, Mande SS.

Author information:

(1)Bio-Sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services

Limited, 1 Software Units Layout, Hyderabad 500 081, Andhra Pradesh, India.

BACKGROUND: Obtaining accurate estimates of microbial diversity using rDNA

profiling is the first step in most metagenomics projects. Consequently, most

metagenomic projects spend considerable amounts of time, money and manpower for

experimentally cloning, amplifying and sequencing the rDNA content in a

metagenomic sample. In the second step, the entire genomic content of the

metagenome is extracted, sequenced and analyzed. Since DNA sequences obtained in

this second step also contain rDNA fragments, rapid in silico identification of

these rDNA fragments would drastically reduce the cost, time and effort of

current metagenomic projects by entirely bypassing the experimental steps of

primer based rDNA amplification, cloning and sequencing. In this study, we

present an algorithm called i-rDNA that can facilitate the rapid detection of 16S

rDNA fragments from amongst millions of sequences in metagenomic data sets with

high detection sensitivity.

RESULTS: Performance evaluation with data sets/database variants simulating

typical metagenomic scenarios indicates the significantly high detection

sensitivity of i-rDNA. Moreover, i-rDNA can process a million sequences in less

than an hour on a simple desktop with modest hardware specifications.

CONCLUSIONS: In addition to the speed of execution, high sensitivity and low

false positive rate, the utility of the algorithmic approach discussed in this

paper is immense given that it would help in bypassing the entire experimental

step of primer-based rDNA amplification, cloning and sequencing. Application of

this algorithmic approach would thus drastically reduce the cost, time and human

efforts invested in all metagenomic projects.

AVAILABILITY: A web-server for the i-rDNA algorithm is available at

http://metagenomics.atc.tcs.com/i-rDNA/

DOI: 10.1186/1471-2164-12-S3-S12

PMCID: PMC3333171

PMID: 22369265 [Indexed for MEDLINE]

1574. BMC Genomics. 2011 Nov 30;12 Suppl 3:S4. doi: 10.1186/1471-2164-12-S3-S4. Epub

2011 Nov 30.

INDUS - a composition-based approach for rapid and accurate taxonomic

classification of metagenomic sequences.

Mohammed MH(1), Ghosh TS, Reddy RM, Reddy CV, Singh NK, Mande SS.

Author information:

(1)Bio-sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services

Limited, 1 Software Units Layout, Madhapur, Hyderabad - 500081, Andhra Pradesh,

India.

BACKGROUND: Taxonomic classification of metagenomic sequences is the first step

in metagenomic analysis. Existing taxonomic classification approaches are of two

types, similarity-based and composition-based. Similarity-based approaches,

though accurate and specific, are extremely slow. Since, metagenomic projects

generate millions of sequences, adopting similarity-based approaches becomes

virtually infeasible for research groups having modest computational resources.

In this study, we present INDUS - a composition-based approach that incorporates

the following novel features. First, INDUS discards the 'one genome-one

composition' model adopted by existing compositional approaches. Second, INDUS

uses 'compositional distance' information for identifying appropriate assignment

levels. Third, INDUS incorporates steps that attempt to reduce biases due to

database representation.

RESULTS: INDUS is able to rapidly classify sequences in both simulated and real

metagenomic sequence data sets with classification efficiency significantly

higher than existing composition-based approaches. Although the classification

efficiency of INDUS is observed to be comparable to those by similarity-based

approaches, the binning time (as compared to alignment based approaches) is 23-33

times lower.

CONCLUSION: Given it's rapid execution time, and high levels of classification

efficiency, INDUS is expected to be of immense interest to researchers working in

metagenomics and microbial ecology.

AVAILABILITY: A web-server for the INDUS algorithm is available at

http://metagenomics.atc.tcs.com/INDUS/

DOI: 10.1186/1471-2164-12-S3-S4

PMCID: PMC3333187

PMID: 22369237 [Indexed for MEDLINE]

1575. BMC Med Genomics. 2011 Nov 28;4:81. doi: 10.1186/1755-8794-4-81.

Estimates of array and pool-construction variance for planning efficient

DNA-pooling genome wide association studies.

Earp MA(1), Rahmani M, Chew K, Brooks-Wilson A.

Author information:

(1)Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver,

BC, Canada.

BACKGROUND: Until recently, genome-wide association studies (GWAS) have been

restricted to research groups with the budget necessary to genotype hundreds, if

not thousands, of samples. Replacing individual genotyping with genotyping of DNA

pools in Phase I of a GWAS has proven successful, and dramatically altered the

financial feasibility of this approach. When conducting a pool-based GWAS, how

well SNP allele frequency is estimated from a DNA pool will influence a study's

power to detect associations. Here we address how to control the variance in

allele frequency estimation when DNAs are pooled, and how to plan and conduct the

most efficient well-powered pool-based GWAS.

METHODS: By examining the variation in allele frequency estimation on SNP arrays

between and within DNA pools we determine how array variance [var(e(array))] and

pool-construction variance [var(e(construction))] contribute to the total

variance of allele frequency estimation. This information is useful in deciding

whether replicate arrays or replicate pools are most useful in reducing variance.

Our analysis is based on 27 DNA pools ranging in size from 74 to 446 individual

samples, genotyped on a collective total of 128 Illumina beadarrays: 24

1M-Single, 32 1M-Duo, and 72 660-Quad.

RESULTS: For all three Illumina SNP array types our estimates of var(e(array))

were similar, between 3-4 × 10-4 for normalized data. Var(e(construction))

accounted for between 20-40% of pooling variance across 27 pools in normalized

data.

CONCLUSIONS: We conclude that relative to var(e(array)), var(e(construction)) is

of less importance in reducing the variance in allele frequency estimation from

DNA pools; however, our data suggests that on average it may be more important

than previously thought. We have prepared a simple online tool, PoolingPlanner

(available at http://www.kchew.ca/PoolingPlanner/), which calculates the

effective sample size (ESS) of a DNA pool given a range of replicate array

values. ESS can be used in a power calculator to perform pool-adjusted

calculations. This allows one to quickly calculate the loss of power associated

with a pooling experiment to make an informed decision on whether a pool-based

GWAS is worth pursuing.

DOI: 10.1186/1755-8794-4-81

PMCID: PMC3247851

PMID: 22122996 [Indexed for MEDLINE]

1576. Database (Oxford). 2011 Nov 26;2011:bar050. doi: 10.1093/database/bar050. Print

2011.

SalmonDB: a bioinformatics resource for Salmo salar and Oncorhynchus mykiss.

Di Génova A(1), Aravena A, Zapata L, González M, Maass A, Iturra P.

Author information:

(1)Laboratory of Bioinformatics and Mathematics of the Genome, Center for

Mathematical Modeling (UMI 2807 CNRS) and Center for Genome Regulation (Fondap

15090007), University of Chile, Santiago, Chile.

SalmonDB is a new multiorganism database containing EST sequences from Salmo

salar, Oncorhynchus mykiss and the whole genome sequence of Danio rerio,

Gasterosteus aculeatus, Tetraodon nigroviridis, Oryzias latipes and Takifugu

rubripes, built with core components from GMOD project, GOPArc system and the

BioMart project. The information provided by this resource includes Gene Ontology

terms, metabolic pathways, SNP prediction, CDS prediction, orthologs prediction,

several precalculated BLAST searches and domains. It also provides a BLAST server

for matching user-provided sequences to any of the databases and an advanced

query tool (BioMart) that allows easy browsing of EST databases with user-defined

criteria. These tools make SalmonDB database a valuable resource for researchers

searching for transcripts and genomic information regarding S. salar and other

salmonid species. The database is expected to grow in the near feature,

particularly with the S. salar genome sequencing project. Database URL:

http://genomicasalmones.dim.uchile.cl/

DOI: 10.1093/database/bar050

PMCID: PMC3225076

PMID: 22120661 [Indexed for MEDLINE]

1577. BMC Bioinformatics. 2011 Nov 15;12:446. doi: 10.1186/1471-2105-12-446.

POPISK: T-cell reactivity prediction using support vector machines and string

kernels.

Tung CW(1), Ziehm M, Kämper A, Kohlbacher O, Ho SY.

Author information:

(1)School of Pharmacy, Kaohsiung Medical University, Kaohsiung 807, Taiwan.

BACKGROUND: Accurate prediction of peptide immunogenicity and characterization of

relation between peptide sequences and peptide immunogenicity will be greatly

helpful for vaccine designs and understanding of the immune system. In contrast

to the prediction of antigen processing and presentation pathway, the prediction

of subsequent T-cell reactivity is a much harder topic. Previous studies of

identifying T-cell receptor (TCR) recognition positions were based on small-scale

analyses using only a few peptides and concluded different recognition positions

such as positions 4, 6 and 8 of peptides with length 9. Large-scale analyses are

necessary to better characterize the effect of peptide sequence variations on

T-cell reactivity and design predictors of a peptide's T-cell reactivity (and

thus immunogenicity). The identification and characterization of important

positions influencing T-cell reactivity will provide insights into the underlying

mechanism of immunogenicity.

RESULTS: This work establishes a large dataset by collecting immunogenicity data

from three major immunology databases. In order to consider the effect of MHC

restriction, peptides are classified by their associated MHC alleles.

Subsequently, a computational method (named POPISK) using support vector machine

with a weighted degree string kernel is proposed to predict T-cell reactivity and

identify important recognition positions. POPISK yields a mean 10-fold

cross-validation accuracy of 68% in predicting T-cell reactivity of

HLA-A2-binding peptides. POPISK is capable of predicting immunogenicity with

scores that can also correctly predict the change in T-cell reactivity related to

point mutations in epitopes reported in previous studies using crystal

structures. Thorough analyses of the prediction results identify the important

positions 4, 6, 8 and 9, and yield insights into the molecular basis for TCR

recognition. Finally, we relate this finding to physicochemical properties and

structural features of the MHC-peptide-TCR interaction.

CONCLUSIONS: A computational method POPISK is proposed to predict immunogenicity

with scores which are useful for predicting immunogenicity changes made by

single-residue modifications. The web server of POPISK is freely available at

http://iclab.life.nctu.edu.tw/POPISK.

DOI: 10.1186/1471-2105-12-446

PMCID: PMC3228774

PMID: 22085524 [Indexed for MEDLINE]

1578. Bioinformatics. 2011 Nov 1;27(21):3074-5. doi: 10.1093/bioinformatics/btr519.

Epub 2011 Sep 11.

Metavir: a web server dedicated to virome analysis.

Roux S(1), Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F.

Author information:

(1)Laboratoire Microorganismes: Génome et Environnement, Clermont Université,

Université Blaise Pascal, BP 10448, F-63000 Clermont-Ferrand, France.

simon.roux@univ-bpclermont.fr

SUMMARY: Metavir is a web server dedicated to the analysis of viral metagenomes

(viromes). In addition to classical approaches for analyzing metagenomes (general

sequence characteristics, taxonomic composition), new tools developed

specifically for viral sequence analysis make it possible to: (i) explore viral

diversity through automatically constructed phylogenies for selected marker

genes, (ii) estimate gene richness through rarefaction curves and (iii) perform

cross-comparison against other viromes using sequence similarities. Metavir is

thus unique as a platform that allows a comprehensive virome analysis.

AVAILABILITY: Metavir is freely available online at:

http://metavir-meb.univ-bpclermont.fr.

CONTACT: simon.roux@univ-bpclermont.fr.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr519

PMID: 21911332 [Indexed for MEDLINE]

1579. Bioinformatics. 2011 Nov 1;27(21):3002-9. doi: 10.1093/bioinformatics/btr513.

Epub 2011 Sep 7.

Protein-protein binding affinity prediction on a diverse set of structures.

Moal IH(1), Agius R, Bates PA.

Author information:

(1)Biomolecular Modelling Laboratory, Cancer Research UK London Research

Institute, London WC2A 3LY, UK.

Comment in

Bioinformatics. 2015 Feb 15;31(4):614-5.

MOTIVATION: Accurate binding free energy functions for protein-protein

interactions are imperative for a wide range of purposes. Their construction is

predicated upon ascertaining the factors that influence binding and their

relative importance. A recent benchmark of binding affinities has allowed, for

the first time, the evaluation and construction of binding free energy models

using a diverse set of complexes, and a systematic assessment of our ability to

model the energetics of conformational changes.

RESULTS: We construct a large set of molecular descriptors using commonly

available tools, introducing the use of energetic factors associated with

conformational changes and disorder to order transitions, as well as features

calculated on structural ensembles. The descriptors are used to train and test a

binding free energy model using a consensus of four machine learning algorithms,

whose performance constitutes a significant improvement over the other state of

the art empirical free energy functions tested. The internal workings of the

learners show how the descriptors are used, illuminating the determinants of

protein-protein binding.

AVAILABILITY: The molecular descriptor set and descriptor values for all

complexes are available in the Supplementary Material. A web server for the

learners and coordinates for the bound and unbound structures can be accessed

from the website: http://bmm.cancerresearchuk.org/~Affinity.

CONTACT: paul.bates@cancer.org.uk.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr513

PMID: 21903632 [Indexed for MEDLINE]

1580. Bioinformatics. 2011 Nov 1;27(21):3010-6. doi: 10.1093/bioinformatics/btr508.

Epub 2011 Sep 6.

SpliceTrap: a method to quantify alternative splicing under single cellular

conditions.

Wu J(1), Akerman M, Sun S, McCombie WR, Krainer AR, Zhang MQ.

Author information:

(1)Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA.

MOTIVATION: Alternative splicing (AS) is a pre-mRNA maturation process leading to

the expression of multiple mRNA variants from the same primary transcript. More

than 90% of human genes are expressed via AS. Therefore, quantifying the

inclusion level of every exon is crucial for generating accurate transcriptomic

maps and studying the regulation of AS.

RESULTS: Here we introduce SpliceTrap, a method to quantify exon inclusion levels

using paired-end RNA-seq data. Unlike other tools, which focus on full-length

transcript isoforms, SpliceTrap approaches the expression-level estimation of

each exon as an independent Bayesian inference problem. In addition, SpliceTrap

can identify major classes of alternative splicing events under a single cellular

condition, without requiring a background set of reads to estimate relative

splicing changes. We tested SpliceTrap both by simulation and real data analysis,

and compared it to state-of-the-art tools for transcript quantification.

SpliceTrap demonstrated improved accuracy, robustness and reliability in

quantifying exon-inclusion ratios.

CONCLUSIONS: SpliceTrap is a useful tool to study alternative splicing

regulation, especially for accurate quantification of local exon-inclusion ratios

from RNA-seq data.

AVAILABILITY AND IMPLEMENTATION: SpliceTrap can be implemented online through the

CSH Galaxy server http://cancan.cshl.edu/splicetrap and is also available for

download and installation at http://rulai.cshl.edu/splicetrap/.

CONTACT: michael.zhang@utdallas.edu.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr508

PMCID: PMC3198574

PMID: 21896509 [Indexed for MEDLINE]

1581. Bioinformatics. 2011 Nov 1;27(21):3076-7. doi: 10.1093/bioinformatics/btr504.

Epub 2011 Sep 6.

wapRNA: a web-based application for the processing of RNA sequences.

Zhao W(1), Liu W, Tian D, Tang B, Wang Y, Yu C, Li R, Ling Y, Wu J, Song S, Hu S.

Author information:

(1)Key Laboratory of Genome Sciences and Information, Beijing Institute of

Genomics, Chinese Academy of Sciences, Beijing 100029, PR China.

SUMMARY: mRNA/miRNA-seq technology is becoming the leading technology to globally

profile gene expression and elucidate the transcriptional regulation mechanisms

in living cells. Although there are many tools available for analyzing RNA-seq

data, few of them are available as easy accessible online web tools for

processing both mRNA and miRNA data for the RNA-seq based user community. As

such, we have developed a comprehensive web application tool for processing

mRNA-seq and miRNA-seq data. Our web tool wapRNA includes four different modules:

mRNA-seq and miRNA-seq sequenced from SOLiD or Solexa platform and all the

modules were tested on previously published experimental data. We accept raw

sequence data with an optional reads filter, followed by mapping and gene

annotation or miRNA prediction. wapRNA also integrates downstream functional

analyses such as Gene Ontology, KEGG pathway, miRNA targets prediction and

comparison of gene's or miRNA's different expression in different samples.

Moreover, we provide the executable packages for installation on user's local

server.

AVAILABILITY: wapRNA is freely available for use at http://waprna.big.ac.cn. The

executable packages and the instruction for installation can be downloaded from

our web site.

CONTACT: husn@big.ac.cn; songshh@big.ac.cn.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr504

PMID: 21896507 [Indexed for MEDLINE]

1582. Brief Bioinform. 2011 Nov;12(6):601-13. doi: 10.1093/bib/bbr050. Epub 2011 Sep 6.

RNA tertiary structure prediction with ModeRNA.

Rother M(1), Rother K, Puton T, Bujnicki JM.

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, 02-109 Warsaw,

Poland.

Noncoding RNAs perform important roles in the cell. As their function is tightly

connected with structure, and as experimental methods are time-consuming and

expensive, the field of RNA structure prediction is developing rapidly. Here, we

present a detailed study on using the ModeRNA software. The tool uses the

comparative modeling approach and can be applied when a structural template is

available and an alignment of reasonable quality can be performed. We guide the

reader through the entire process of modeling Escherichia coli tRNA(Thr) in a

conformation corresponding to the complex with an aminoacyl-tRNA synthetase

(aaRS). We describe the choice of a template structure, preparation of input

files, and explore three possible modeling strategies. In the end, we evaluate

the resulting models using six alternative benchmarks. The ModeRNA software can

be freely downloaded from http://iimcb.genesilico.pl/moderna/ under the

conditions of the General Public License. It runs under LINUX, Windows and Mac

OS. It is also available as a server at

http://iimcb.genesilico.pl/modernaserver/. The models and the script to reproduce

the study from this article are available at

http://www.genesilico.pl/moderna/examples/.

DOI: 10.1093/bib/bbr050

PMID: 21896613 [Indexed for MEDLINE]

1583. J Mol Graph Model. 2011 Nov;31:28-34. doi: 10.1016/j.jmgm.2011.08.001. Epub 2011

Aug 11.

A novel method for quantitatively predicting non-covalent interactions from

protein and nucleic acid sequence.

Wu J(1), Hu D, Xu X, Ding Y, Yan S, Sun X.

Author information:

(1)School of Geography and Biological Information, Nanjing University of Posts

and Telecommunications, Nanjing, PR China.

Biochemical interactions between proteins and biological macromolecules are

dominated by non-covalent interactions. A novel method is presented for

quantitatively predicting the number of two most dominant non-covalent

interactions, i.e., hydrogen bonds and van der Waals contacts, potentially

forming in a hypothetical protein-nucleic acid complex from sequences using

support vector machine regression models in conjunction with a hybrid feature.

The hybrid feature consists of the sequence-length fraction information, conjoint

triad for protein sequences and the gapped dinucleotide composition. The

SVR-based models achieved excellent performance. The polarity of amino acids was

also found to play a vital role in the formation of hydrogen bonds and van der

Waals contacts. We have constructed a web server H-VDW

(http://www.cbi.seu.edu.cn/H-VDW/H-VDW.htm) for public access to the SVR models.

Copyright © 2011 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.jmgm.2011.08.001

PMID: 21920789 [Indexed for MEDLINE]

1584. Mol Cell Proteomics. 2011 Nov;10(11):M111.010629. doi: 10.1074/mcp.M111.010629.

Epub 2011 Aug 11.

Large-scale de novo prediction of physical protein-protein association.

Elefsinioti A(1), Saraç ÖS, Hegele A, Plake C, Hubner NC, Poser I, Sarov M, Hyman

A, Mann M, Schroeder M, Stelzl U, Beyer A.

Author information:

(1)Biotechnology Center, TU Dresden, Dresden, Germany.

Information about the physical association of proteins is extensively used for

studying cellular processes and disease mechanisms. However, complete

experimental mapping of the human interactome will remain prohibitively difficult

in the near future. Here we present a map of predicted human protein interactions

that distinguishes functional association from physical binding. Our network

classifies more than 5 million protein pairs predicting 94,009 new interactions

with high confidence. We experimentally tested a subset of these predictions

using yeast two-hybrid analysis and affinity purification followed by

quantitative mass spectrometry. Thus we identified 462 new protein-protein

interactions and confirmed the predictive power of the network. These independent

experiments address potential issues of circular reasoning and are a distinctive

feature of this work. Analysis of the physical interactome unravels subnetworks

mediating between different functional and physical subunits of the cell.

Finally, we demonstrate the utility of the network for the analysis of molecular

mechanisms of complex diseases by applying it to genome-wide association studies

of neurodegenerative diseases. This analysis provides new evidence implying

TOMM40 as a factor involved in Alzheimer's disease. The network provides a

high-quality resource for the analysis of genomic data sets and genetic

association studies in particular. Our interactome is available via the hPRINT

web server at: www.print-db.org.

DOI: 10.1074/mcp.M111.010629

PMCID: PMC3226409

PMID: 21836163 [Indexed for MEDLINE]

1585. RNA Biol. 2011 Nov-Dec;8(6):988-96. doi: 10.4161/rna.8.6.17813. Epub 2011 Nov 1.

Highly accurate and high-resolution function prediction of RNA binding proteins

by fold recognition and binding affinity prediction.

Zhao H(1), Yang Y, Zhou Y.

Author information:

(1)School of Informatics, Indiana University Purdue University, Indianapolis, IN,

USA.

A full understanding of the mechanism of post- transcriptional regulation

requires more than simple two- state prediction (binding or not binding) for RNA

binding proteins. Here we report a sequence-based technique dedicated for

predicting complex structures of protein and RNA by combining fold recognition

with binding affinity prediction. The method not only provides a highly accurate

complex structure prediction (77% of residues are within 4°A RMSD from native in

average for the independent test set) but also achieves the best performing

two-state binding or non-binding prediction with an accuracy of 98%, precision of

84%, and Mathews correlation coefficient (MCC) of 0.62. Moreover, it predicts

binding residues with an accuracy of 84%, precision of 66% and MCC value of 0.51.

In addition, it has a success rate of 77% in predicting RNA binding types (mRNA,

tRNA or rRNA). We further demonstrate that it makes more than 10% improvement

either in precision or sensitivity than PSI- BLAST, HHPRED and our previously

developed structure- based technique. This method expects to be useful for highly

accurate genome-scale, high-resolution prediction of RNA-binding proteins and

their complex structures. A web server (SPOT) is freely available for academic

users at http://sparks.informatics.iupui.edu.

DOI: 10.4161/rna.8.6.17813

PMCID: PMC3360076

PMID: 21955494 [Indexed for MEDLINE]

1586. J Biomed Semantics. 2011 Oct 31;2(1):9. doi: 10.1186/2041-1480-2-9.

Brucellosis Ontology (IDOBRU) as an extension of the Infectious Disease Ontology.

Lin Y(1), Xiang Z, He Y.

Author information:

(1)Unit for Laboratory Animal Medicine, University of Michigan Medical School,

Ann Arbor, MI 48109, USA. yongqunh@umich.edu.

BACKGROUND: Caused by intracellular Gram-negative bacteria Brucella spp.,

brucellosis is the most common bacterial zoonotic disease. Extensive studies in

brucellosis have yielded a large number of publications and data covering various

topics ranging from basic Brucella genetic study to vaccine clinical trials. To

support data interoperability and reasoning, a community-based

brucellosis-specific biomedical ontology is needed.

RESULTS: The Brucellosis Ontology (IDOBRU:

http://sourceforge.net/projects/idobru), a biomedical ontology in the brucellosis

domain, is an extension ontology of the core Infectious Disease Ontology

(IDO-core) and follows OBO Foundry principles. Currently IDOBRU contains 1503

ontology terms, which includes 739 Brucella-specific terms, 414 IDO-core terms,

and 350 terms imported from 10 existing ontologies. IDOBRU has been used to model

different aspects of brucellosis, including host infection, zoonotic disease

transmission, symptoms, virulence factors and pathogenesis, diagnosis,

intentional release, vaccine prevention, and treatment. Case studies are

typically used in our IDOBRU modeling. For example, diurnal temperature variation

in Brucella patients, a Brucella-specific PCR method, and a WHO-recommended

brucellosis treatment were selected as use cases to model brucellosis symptom,

diagnosis, and treatment, respectively. Developed using OWL, IDOBRU supports

OWL-based ontological reasoning. For example, by performing a Description Logic

(DL) query in the OWL editor Protégé 4 or a SPARQL query in an IDOBRU SPARQL

server, a check of Brucella virulence factors showed that eight of them are known

protective antigens based on the biological knowledge captured within the

ontology.

CONCLUSIONS: IDOBRU is the first reported bacterial infectious disease ontology

developed to represent different disease aspects in a formal logical format. It

serves as a brucellosis knowledgebase and supports brucellosis data integration

and automated reasoning.

DOI: 10.1186/2041-1480-2-9

PMCID: PMC3217896

PMID: 22041276

1587. Nurs Manag (Harrow). 2011 Oct 26;18(7):14.

Patient-based improvement.

Williams R.

The King's Fund has produced an online tool to help healthcare staff create

patient-based improvement initiatives. Experience-based co-design (EBCD) enables

staff and service users to design services or care pathways in partnership.

Details are available at www.kingsfund.org.uk/ebcd/index.html.

DOI: 10.7748/nm.18.7.14.s8

PMID: 27741602

1588. Bioinformatics. 2011 Oct 15;27(20):2812-9. doi: 10.1093/bioinformatics/btr494.

Epub 2011 Aug 27.

SCLpred: protein subcellular localization prediction by N-to-1 neural networks.

Mooney C(1), Wang YH, Pollastri G.

Author information:

(1)School of Computer Science and Informatics, University College Dublin,

Belfield, Ireland.

SUMMARY: Knowledge of the subcellular location of a protein provides valuable

information about its function and possible interaction with other proteins. In

the post-genomic era, fast and accurate predictors of subcellular location are

required if this abundance of sequence data is to be fully exploited. We have

developed a subcellular localization predictor (SCLpred), which predicts the

location of a protein into four classes for animals and fungi and five classes

for plants (secreted, cytoplasm, nucleus, mitochondrion and chloroplast) using

machine learning models trained on large non-redundant sets of protein sequences.

The algorithm powering SCLpred is a novel Neural Network (N-to-1 Neural Network,

or N1-NN) we have developed, which is capable of mapping whole sequences into

single properties (a functional class, in this work) without resorting to

predefined transformations, but rather by adaptively compressing the sequence

into a hidden feature vector. We benchmark SCLpred against other publicly

available predictors using two benchmarks including a new subset of Swiss-Prot

Release 2010\_06. We show that SCLpred surpasses the state of the art. The N1-NN

algorithm is fully general and may be applied to a host of problems of similar

shape, that is, in which a whole sequence needs to be mapped into a fixed-size

array of properties, and the adaptive compression it operates may shed light on

the space of protein sequences.

AVAILABILITY: The predictive systems described in this article are publicly

available as a web server at http://distill.ucd.ie/distill/.

CONTACT: gianluca.pollastri@ucd.ie.

DOI: 10.1093/bioinformatics/btr494

PMID: 21873639 [Indexed for MEDLINE]

1589. BMC Genomics. 2011 Oct 14;12:507. doi: 10.1186/1471-2164-12-507.

Chipster: user-friendly analysis software for microarray and other

high-throughput data.

Kallio MA(1), Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M,

Käki J, Korpelainen EI.

Author information:

(1)CSC - IT Center for Science, Keilaranta 14, Keilaniemi, Espoo, Finland.

BACKGROUND: The growth of high-throughput technologies such as microarrays and

next generation sequencing has been accompanied by active research in data

analysis methodology, producing new analysis methods at a rapid pace. While most

of the newly developed methods are freely available, their use requires

substantial computational skills. In order to enable non-programming biologists

to benefit from the method development in a timely manner, we have created the

Chipster software.

RESULTS: Chipster (http://chipster.csc.fi/) brings a powerful collection of data

analysis methods within the reach of bioscientists via its intuitive graphical

user interface. Users can analyze and integrate different data types such as gene

expression, miRNA and aCGH. The analysis functionality is complemented with rich

interactive visualizations, allowing users to select datapoints and create new

gene lists based on these selections. Importantly, users can save the performed

analysis steps as reusable, automatic workflows, which can also be shared with

other users. Being a versatile and easily extendable platform, Chipster can be

used for microarray, proteomics and sequencing data. In this article we describe

its comprehensive collection of analysis and visualization tools for microarray

data using three case studies.

CONCLUSIONS: Chipster is a user-friendly analysis software for high-throughput

data. Its intuitive graphical user interface enables biologists to access a

powerful collection of data analysis and integration tools, and to visualize data

interactively. Users can collaborate by sharing analysis sessions and workflows.

Chipster is open source, and the server installation package is freely available.

DOI: 10.1186/1471-2164-12-507

PMCID: PMC3215701

PMID: 21999641 [Indexed for MEDLINE]

1590. BMC Bioinformatics. 2011 Oct 5;12:387. doi: 10.1186/1471-2105-12-387.

Biblio-MetReS: a bibliometric network reconstruction application and server.

Usié A(1), Karathia H, Teixidó I, Valls J, Faus X, Alves R, Solsona F.

Author information:

(1)Department d'Informàtica i Enginyeria Industrial, Universitat de Lleida, Av,

Jaume II n°69, 25001 Lleida, Spain.

BACKGROUND: Reconstruction of genes and/or protein networks from automated

analysis of the literature is one of the current targets of text mining in

biomedical research. Some user-friendly tools already perform this analysis on

precompiled databases of abstracts of scientific papers. Other tools allow expert

users to elaborate and analyze the full content of a corpus of scientific

documents. However, to our knowledge, no user friendly tool that simultaneously

analyzes the latest set of scientific documents available on line and

reconstructs the set of genes referenced in those documents is available.

RESULTS: This article presents such a tool, Biblio-MetReS, and compares its

functioning and results to those of other user-friendly applications (iHOP,

STRING) that are widely used. Under similar conditions, Biblio-MetReS creates

networks that are comparable to those of other user friendly tools. Furthermore,

analysis of full text documents provides more complete reconstructions than those

that result from using only the abstract of the document.

CONCLUSIONS: Literature-based automated network reconstruction is still far from

providing complete reconstructions of molecular networks. However, its value as

an auxiliary tool is high and it will increase as standards for reporting

biological entities and relationships become more widely accepted and enforced.

Biblio-MetReS is an application that can be downloaded from

http://metres.udl.cat/. It provides an easy to use environment for researchers to

reconstruct their networks of interest from an always up to date set of

scientific documents.

DOI: 10.1186/1471-2105-12-387

PMCID: PMC3228545

PMID: 21975133 [Indexed for MEDLINE]

1591. BMC Genomics. 2011 Oct 4;12:483. doi: 10.1186/1471-2164-12-483.

Ortho2ExpressMatrix--a web server that interprets cross-species gene expression

data by gene family information.

Meinel T(1), Schweiger MR, Ludewig AH, Chenna R, Krobitsch S, Herwig R.

Author information:

(1)Structural Bioinformatics Group, Institute for Physiology, Charité-University

Medicine Berlin, Thielallee 71, 14195 Berlin, Germany. thomas.meinel@charite.de

BACKGROUND: The study of gene families is pivotal for the understanding of gene

evolution across different organisms and such phylogenetic background is often

used to infer biochemical functions of genes. Modern high-throughput experiments

offer the possibility to analyze the entire transcriptome of an organism;

however, it is often difficult to deduct functional information from that data.

RESULTS: To improve functional interpretation of gene expression we introduce

Ortho2ExpressMatrix, a novel tool that integrates complex gene family

information, computed from sequence similarity, with comparative gene expression

profiles of two pre-selected biological objects: gene families are displayed with

two-dimensional matrices. Parameters of the tool are object type (two organisms,

two individuals, two tissues, etc.), type of computational gene family inference,

experimental meta-data, microarray platform, gene annotation level and genome

build. Family information in Ortho2ExpressMatrix bases on computationally

different protein family approaches such as EnsemblCompara, InParanoid, SYSTERS

and Ensembl Family. Currently, respective all-against-all associations are

available for five species: human, mouse, worm, fruit fly and yeast.

Additionally, microRNA expression can be examined with respect to miRBase or

TargetScan families. The visualization, which is typical for Ortho2ExpressMatrix,

is performed as matrix view that displays functional traits of genes

(differential expression) as well as sequence similarity of protein family

members (BLAST e-values) in colour codes. Such translations are intended to

facilitate the user's perception of the research object.

CONCLUSIONS: Ortho2ExpressMatrix integrates gene family information with

genome-wide expression data in order to enhance functional interpretation of

high-throughput analyses on diseases, environmental factors, or genetic

modification or compound treatment experiments. The tool explores differential

gene expression in the light of orthology, paralogy and structure of gene

families up to the point of ambiguity analyses. Results can be used for filtering

and prioritization in functional genomic, biomedical and systems biology

applications. The web server is freely accessible at

http://bioinf-data.charite.de/o2em/cgi-bin/o2em.pl.

DOI: 10.1186/1471-2164-12-483

PMCID: PMC3202273

PMID: 21970648 [Indexed for MEDLINE]

1592. Bioinformatics. 2011 Oct 1;27(19):2758-60. doi: 10.1093/bioinformatics/btr435.

Epub 2011 Jul 26.

ReplacementMatrix: a web server for maximum-likelihood estimation of amino acid

replacement rate matrices.

Dang CC(1), Lefort V, Le VS, Le QS, Gascuel O.

Author information:

(1)College of Technology and Information Technology Institute, Vietnam National

University, Hanoi, Vietnam.

SUMMARY: Amino acid replacement rate matrices are an essential basis of protein

studies (e.g. in phylogenetics and alignment). A number of general purpose

matrices have been proposed (e.g. JTT, WAG, LG) since the seminal work of

Margaret Dayhoff and co-workers. However, it has been shown that matrices

specific to certain protein groups (e.g. mitochondrial) or life domains (e.g.

viruses) differ significantly from general average matrices, and thus perform

better when applied to the data to which they are dedicated. This Web server

implements the maximum-likelihood estimation procedure that was used to estimate

LG, and provides a number of tools and facilities. Users upload a set of multiple

protein alignments from their domain of interest and receive the resulting matrix

by email, along with statistics and comparisons with other matrices. A

non-parametric bootstrap is performed optionally to assess the variability of

replacement rate estimates. Maximum-likelihood trees, inferred using the

estimated rate matrix, are also computed optionally for each input alignment.

Finely tuned procedures and up-to-date ML software (PhyML 3.0, XRATE) are

combined to perform all these heavy calculations on our clusters.

AVAILABILITY: http://www.atgc-montpellier.fr/ReplacementMatrix/

CONTACT: olivier.gascuel@lirmm.fr

SUPPLEMENTARY INFORMATION: Supplementary data are available at

http://www.atgc-montpellier.fr/ReplacementMatrix/

DOI: 10.1093/bioinformatics/btr435

PMID: 21791535 [Indexed for MEDLINE]

1593. Comput Biol Med. 2011 Oct;41(10):946-59. doi: 10.1016/j.compbiomed.2011.08.005.

Epub 2011 Aug 30.

Improving protein secondary structure prediction using a multi-modal BP method.

Qu W(1), Sui H, Yang B, Qian W.

Author information:

(1)School of Information Engineering, University of Science and Technology

Beijing, China. quwu.ustb@gmail.com

Methods for predicting protein secondary structures provide information that is

useful both in ab initio structure prediction and as additional restraints for

fold recognition algorithms. Secondary structure predictions may also be used to

guide the design of site directed mutagenesis studies, and to locate potential

functionally important residues. In this article, we propose a multi-modal back

propagation neural network (MMBP) method for predicting protein secondary

structures. Using a Knowledge Discovery Theory based on Inner Cognitive Mechanism

(KDTICM) method, we have constructed a compound pyramid model (CPM), which is

composed of three layers of intelligent interface that integrate multi-modal back

propagation neural network (MMBP), mixed-modal SVM (MMS), modified Knowledge

Discovery in Databases (KDD(⁎)) process and so on. The CPM method is both an

integrated web server and a standalone application that exploits recent

advancements in knowledge discovery and machine learning to perform very accurate

protein secondary structure predictions. Using a non-redundant test dataset of

256 proteins from RCASP256, the CPM method achieves an average Q(3) score of

86.13% (SOV99=84.66%). Extensive testing indicates that this is significantly

better than any other method currently available. Assessments using RS126 and

CB513 datasets indicate that the CPM method can achieve average Q(3) score

approaching 83.99% (SOV99=80.25%) and 85.58% (SOV99=81.15%). By using both

sequence and structure databases and by exploiting the latest techniques in

machine learning it is possible to routinely predict protein secondary structure

with an accuracy well above 80%. A program and web server, called CPM, which

performs these secondary structure predictions, is accessible at

http://kdd.ustb.edu.cn/protein\_Web/.

Copyright © 2011 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiomed.2011.08.005

PMID: 21880310 [Indexed for MEDLINE]

1594. Genomics Proteomics Bioinformatics. 2011 Oct;9(4-5):179-82. doi:

10.1016/S1672-0229(11)60021-1.

Junker: an intergenic explorer for bacterial genomes.

Sridhar J(1), Sabarinathan R, Balan SS, Rafi ZA, Gunasekaran P, Sekar K.

Author information:

(1)Centre of Excellence in Bioinformatics, School of Biotechnology, Madurai

Kamaraj University, Madurai 625021, Tamilnadu, India.

In the past few decades, scientists from all over the world have taken a keen

interest in novel functional units such as small regulatory RNAs, small open

reading frames, pseudogenes, transposons, integrase binding attB/attP sites,

repeat elements within the bacterial intergenic regions (IGRs) and in the

analysis of those "junk" regions for genomic complexity. Here we have developed a

web server, named Junker, to facilitate the in-depth analysis of IGRs for

examining their length distribution, four-quadrant plots, GC percentage and

repeat details. Upon selection of a particular bacterial genome, the physical

genome map is displayed as a multiple loci with options to view any loci of

interest in detail. In addition, an IGR statistics module has been created and

implemented in the web server to analyze the length distribution of the IGRs and

to understand the disordered grouping of IGRs across the genome by generating the

four-quadrant plots. The proposed web server is freely available at the URL

http://pranag.physics.iisc.ernet.in/junker/.

Copyright © 2011 Beijing Genomics Institute. Published by Elsevier Ltd. All

rights reserved.

DOI: 10.1016/S1672-0229(11)60021-1

PMCID: PMC5054447

PMID: 22196361 [Indexed for MEDLINE]

1595. Genomics Proteomics Bioinformatics. 2011 Oct;9(4-5):171-8. doi:

10.1016/S1672-0229(11)60020-X.

Mining genomic patterns in Mycobacterium tuberculosis H37Rv using a web server

Tuber-Gene.

Rishishwar L(1), Pant B, Pant K, Pardasani KR.

Author information:

(1)School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332,

USA. lavanya.rishishwar@gatech.edu

Mycobacterium tuberculosis (MTB), causative agent of tuberculosis, is one of the

most dreaded diseases of the century. It has long been studied by researchers

throughout the world using various wet-lab and dry-lab techniques. In this study,

we focus on mining useful patterns at genomic level that can be applied for in

silico functional characterization of genes from the MTB complex. The model

developed on the basis of the patterns found in this study can correctly identify

99.77% of the input genes from the genome of MTB strain H37Rv. The model was

tested against four other MTB strains and the homologue M. bovis to further

evaluate its generalization capability. The mean prediction accuracy was 85.76%.

It was also observed that the GC content remained fairly constant throughout the

genome, implicating the absence of any pathogenicity island transferred from

other organisms. This study reveals that dinucleotide composition is an efficient

functional class discriminator for MTB complex. To facilitate the application of

this model, a web server Tuber-Gene has been developed, which can be freely

accessed at http://www.bifmanit.org/tb2/.

Copyright © 2011 Beijing Genomics Institute. Published by Elsevier Ltd. All

rights reserved.

DOI: 10.1016/S1672-0229(11)60020-X

PMCID: PMC5054438

PMID: 22196360 [Indexed for MEDLINE]

1596. Protein Pept Lett. 2011 Oct;18(10):1010-20.

3dswap-pred: prediction of 3D domain swapping from protein sequence using Random

Forest approach.

Shameer K(1), Pugalenthi G, Kandaswamy KK, Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences, UAS-GKVK Campus, Bellary Road,

Bangalore 560 065, India.

3D domain swapping is a protein structural phenomenon that mediates the formation

of the higher order oligomers in a variety of proteins with different structural

and functional properties. 3D domain swapping is associated with a variety of

biological functions ranging from oligomerization to pathological conformational

diseases. 3D domain swapping is realised subsequent to structure determination

where the protein is observed in the swapped conformation in the oligomeric

state. This is a limiting step to understand this important structural phenomenon

in a large scale from the growing sequence data. A new machine learning approach,

3dswap-pred, has been developed for the prediction of 3D domain swapping in

protein structures from mere sequence data using the Random Forest approach.

3Dswap-pred is implemented using a positive sequence dataset derived from

literature based structural curation of 297 structures. A negative sequence

dataset is obtained from 462 SCOP domains using a new sequence data mining

approach and a set of 126 sequencederived features. Statistical validation using

an independent dataset of 68 positive sequences and 313 negative sequences

revealed that 3dswap-pred achieved an accuracy of 63.8%. A webserver is also

implemented using the 3dswap-pred Random Forest model. The server is available

from the URL: http://caps.ncbs.res.in/3dswap-pred.

PMID: 21592079 [Indexed for MEDLINE]

1597. BMC Bioinformatics. 2011 Sep 30;12:385. doi: 10.1186/1471-2105-12-385.

Interactive metagenomic visualization in a Web browser.

Ondov BD(1), Bergman NH, Phillippy AM.

Author information:

(1)National Biodefense Analysis and Countermeasures Center, 110 Thomas Johnson

Drive, Frederick, MD 21702, USA. ondovb@nbacc.net

BACKGROUND: A critical output of metagenomic studies is the estimation of

abundances of taxonomical or functional groups. The inherent uncertainty in

assignments to these groups makes it important to consider both their

hierarchical contexts and their prediction confidence. The current tools for

visualizing metagenomic data, however, omit or distort quantitative hierarchical

relationships and lack the facility for displaying secondary variables.

RESULTS: Here we present Krona, a new visualization tool that allows intuitive

exploration of relative abundances and confidences within the complex hierarchies

of metagenomic classifications. Krona combines a variant of radial, space-filling

displays with parametric coloring and interactive polar-coordinate zooming. The

HTML5 and JavaScript implementation enables fully interactive charts that can be

explored with any modern Web browser, without the need for installed software or

plug-ins. This Web-based architecture also allows each chart to be an independent

document, making them easy to share via e-mail or post to a standard Web server.

To illustrate Krona's utility, we describe its application to various metagenomic

data sets and its compatibility with popular metagenomic analysis tools.

CONCLUSIONS: Krona is both a powerful metagenomic visualization tool and a

demonstration of the potential of HTML5 for highly accessible bioinformatic

visualizations. Its rich and interactive displays facilitate more informed

interpretations of metagenomic analyses, while its implementation as a

browser-based application makes it extremely portable and easily adopted into

existing analysis packages. Both the Krona rendering code and conversion tools

are freely available under a BSD open-source license, and available from:

http://krona.sourceforge.net.

DOI: 10.1186/1471-2105-12-385

PMCID: PMC3190407

PMID: 21961884 [Indexed for MEDLINE]

1598. BMC Bioinformatics. 2011 Sep 26;12:379. doi: 10.1186/1471-2105-12-379.

GOmotif: A web server for investigating the biological role of protein sequence

motifs.

Bristow F(1), He R, Van Domselaar G.

Author information:

(1)National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg,

MB, R3E 3R2, Canada.

BACKGROUND: Many proteins contain conserved sequence patterns (motifs) that

contribute to their functionality. The process of experimentally identifying and

validating novel protein motifs can be difficult, expensive, and time consuming.

A means for helping to identify in advance the possible function of a novel motif

is important to test hypotheses concerning the biological relevance of these

motifs, thus reducing experimental trial-and-error.

RESULTS: GOmotif accepts PROSITE and regular expression formatted motifs as input

and searches a Gene Ontology annotated protein database using motif search tools.

The search returns the set of proteins containing matching motifs and their

associated Gene Ontology terms. These results are presented as: 1) a

hierarchical, navigable tree separated into the three Gene Ontology biological

domains - biological process, cellular component, and molecular function; 2)

corresponding pie charts indicating raw and statistically adjusted distributions

of the results, and 3) an interactive graphical network view depicting the

location of the results in the Gene Ontology.

CONCLUSIONS: GOmotif is a web-based tool designed to assist researchers in

investigating the biological role of novel protein motifs. GOmotif can be freely

accessed at http://www.gomotif.ca.

DOI: 10.1186/1471-2105-12-379

PMCID: PMC3228539

PMID: 21943350 [Indexed for MEDLINE]

1599. Bioinformatics. 2011 Sep 15;27(18):2595-7. doi: 10.1093/bioinformatics/btr440.

Epub 2011 Jul 26.

VISTA Region Viewer (RViewer)--a computational system for prioritizing genomic

intervals for biomedical studies.

Lukashin I(1), Novichkov P, Boffelli D, Paciorkowski AR, Minovitsky S, Yang S,

Dubchak I.

Author information:

(1)Genomics Division, Lawrence Berkeley National Laboratory, MS 84-171, Berkeley,

CA 94720, USA.

SUMMARY: Current genome browsers are designed for linear browsing of individual

genomic regions, but the high-throughput nature of experiments aiming to

elucidate the genetic component of human disease makes it very important to

develop user-friendly tools for comparing several genomic regions in parallel and

prioritizing them based on their functional content. We introduce VISTA Region

Viewer (RViewer), an interactive online tool that allows for efficient screening

and prioritization of regions of the human genome for follow-up studies. The tool

takes as input genetic variation data from different biomedical studies,

determines a number of various functional parameters for both coding and

non-coding sequences in each region and allows for sorting and searching the

results of the analysis in multiple ways.

AVAILABILITY AND IMPLEMENTATION: The tool is implemented as a web application and

is freely accessible on the Web at http://rviewer.lbl.gov

CONTACT: rviewer@lbl.gov; ildubchak@lbl.gov.

DOI: 10.1093/bioinformatics/btr440

PMCID: PMC3167054

PMID: 21791533 [Indexed for MEDLINE]

1600. BMC Genomics. 2011 Sep 7;12:444. doi: 10.1186/1471-2164-12-444.

WebMGA: a customizable web server for fast metagenomic sequence analysis.

Wu S(1), Zhu Z, Fu L, Niu B, Li W.

Author information:

(1)Center for Research in Biological Systems, University of California San Diego,

La Jolla, California 92093, USA.

BACKGROUND: The new field of metagenomics studies microorganism communities by

culture-independent sequencing. With the advances in next-generation sequencing

techniques, researchers are facing tremendous challenges in metagenomic data

analysis due to huge quantity and high complexity of sequence data. Analyzing

large datasets is extremely time-consuming; also metagenomic annotation involves

a wide range of computational tools, which are difficult to be installed and

maintained by common users. The tools provided by the few available web servers

are also limited and have various constraints such as login requirement, long

waiting time, inability to configure pipelines etc.

RESULTS: We developed WebMGA, a customizable web server for fast metagenomic

analysis. WebMGA includes over 20 commonly used tools such as ORF calling,

sequence clustering, quality control of raw reads, removal of sequencing

artifacts and contaminations, taxonomic analysis, functional annotation etc.

WebMGA provides users with rapid metagenomic data analysis using fast and

effective tools, which have been implemented to run in parallel on our local

computer cluster. Users can access WebMGA through web browsers or programming

scripts to perform individual analysis or to configure and run customized

pipelines. WebMGA is freely available at

http://weizhongli-lab.org/metagenomic-analysis.

CONCLUSIONS: WebMGA offers to researchers many fast and unique tools and great

flexibility for complex metagenomic data analysis.

DOI: 10.1186/1471-2164-12-444

PMCID: PMC3180703

PMID: 21899761 [Indexed for MEDLINE]

1601. J Theor Biol. 2011 Sep 7;284(1):42-51. doi: 10.1016/j.jtbi.2011.06.005. Epub 2011

Jun 17.

iLoc-Virus: a multi-label learning classifier for identifying the subcellular

localization of virus proteins with both single and multiple sites.

Xiao X(1), Wu ZC, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China.

xxiao@gordonlifescience.org

In the last two decades or so, although many computational methods were developed

for predicting the subcellular locations of proteins according to their sequence

information, it is still remains as a challenging problem, particularly when the

system concerned contains both single- and multiple-location proteins. Also,

among the existing methods, very few were developed specialized for dealing with

viral proteins, those generated by viruses. Actually, knowledge of the

subcellular localization of viral proteins in a host cell or virus-infected cell

is very important because it is closely related to their destructive tendencies

and consequences. In this paper, by introducing the "multi-label scale" and by

hybridizing the gene ontology information with the sequential evolution

information, a predictor called iLoc-Virus is developed. It can be utilized to

identify viral proteins among the following six locations: (1) viral capsid, (2)

host cell membrane, (3) host endoplasmic reticulum, (4) host cytoplasm, (5) host

nucleus, and (6) secreted. The iLoc-Virus predictor not only can more accurately

predict the location sites of viral proteins in a host cell, but also have the

capacity to deal with virus proteins having more than one location. As a

user-friendly web-server, iLoc-Virus is freely accessible to the public at

http://icpr.jci.edu.cn/bioinfo/iLoc-Virus. Meanwhile, a step-by-step guide is

provided on how to use the web-server to get the desired results. Furthermore,

for the user's convenience, the iLoc-Virus web-server also has the function to

accept the batch job submission. It is anticipated that iLoc-Virus may become a

useful high throughput tool for both basic research and drug development.

Copyright © 2011 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2011.06.005

PMID: 21684290 [Indexed for MEDLINE]

1602. Acta Pharmacol Sin. 2011 Sep;32(9):1116-27. doi: 10.1038/aps.2011.86. Epub 2011

Aug 15.

Proteome reference map and regulation network of neonatal rat cardiomyocyte.

Li ZJ(1), Liu N, Han QD, Zhang YY.

Author information:

(1)Institute of Vascular Medicine, Peking University Third Hospital and Key

Laboratory of Cardiovascular Molecular Biology and Regulatory Peptides, Ministry

of Health and Key Laboratory of Molecular Cardiology, Ministry of Education,

Beijing 100191, China.

AIM: To study and establish a proteome reference map and regulation network of

neonatal rat cardiomyocyte.

METHODS: Cultured cardiomyocytes of neonatal rats were used. All proteins

expressed in the cardiomyocytes were separated and identified by two-dimensional

polyacrylamide gel electrophoresis (2-DE) and matrix-assisted laser

desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS). Biological

networks and pathways of the neonatal rat cardiomyocytes were analyzed using the

Ingenuity Pathway Analysis (IPA) program (www.ingenuity.com). A 2-DE database was

made accessible on-line by Make2ddb package on a web server.

RESULTS: More than 1000 proteins were separated on 2D gels, and 148 proteins were

identified. The identified proteins were used for the construction of an

extensible markup language-based database. Biological networks and pathways were

constructed to analyze the functions associate with cardiomyocyte proteins in the

database. The 2-DE database of rat cardiomyocyte proteins can be accessed at

http://2d.bjmu.edu.cn.

CONCLUSION: A proteome reference map and regulation network of the neonatal rat

cardiomyocytes have been established, which may serve as an international

platform for storage, analysis and visualization of cardiomyocyte proteomic data.

DOI: 10.1038/aps.2011.86

PMCID: PMC4003303

PMID: 21841810 [Indexed for MEDLINE]

1603. Bioeng Bugs. 2011 Sep-Oct;2(5):284-7. doi: 10.4161/bbug.2.5.16113. Epub 2011 Sep

1.

MrBac: a web server for draft metabolic network reconstructions for bacteria.

Liao YC(1), Chen JC, Tsai MH, Tang YH, Chen FC, Hsiung CA.

Author information:

(1)Division of Biostatistics and Bioinformatics, Institute of Population Health

Sciences, National Health Research Institutes, Zhunan 350, Taiwan.

Genome-scale metabolic network reconstruction can be used for simulating cellular

behaviors by simultaneously monitoring thousands of biochemical reactions, and is

therefore important for systems biology studies in microbes. However, the

labor-intensive and time-consuming reconstruction process has hindered the

progress of this important field. Here we present a web server, MrBac (Metabolic

network Reconstructions for Bacteria), to streamline the network reconstruction

process for draft genome-scale metabolic networks and to provide annotation

information from multiple databases for further curation of the draft

reconstructions. MrBac integrates comparative genomics, retrieval of genome

annotations, and generation of standard systems biology file format ready for

network analyses. We also used MrBac to automatically generate a draft metabolic

model of Salmonella enteric serovar Typhimurium LT2. The high similarity between

this automatic model and the experimentally validated models further supports the

usefulness and accuracy of MrBac. The high efficiency and accuracy of MrBac may

accelerate the advances of systems biology studies on microbiology. MrBac is

freely available at http://sb.nhri.org.tw/MrBac.

DOI: 10.4161/bbug.2.5.16113

PMID: 22008641 [Indexed for MEDLINE]

1604. Bioinformatics. 2011 Sep 1;27(17):2384-90. doi: 10.1093/bioinformatics/btr415.

Epub 2011 Jul 14.

A Grid-enabled web portal for NMR structure refinement with AMBER.

Bertini I(1), Case DA, Ferella L, Giachetti A, Rosato A.

Author information:

(1)Magnetic Resonance Center (CERM), University of Florence, Via L. Sacconi 6,

Italy. ivanobertini@cerm.unifi.it

MOTIVATION: The typical workflow for NMR structure determination involves

collecting thousands of conformational restraints, calculating a bundle of 20-40

conformers in agreement with them and refining the energetics of these

conformers. The structure calculation step employs simulated annealing based on

molecular dynamics (MD) simulations with very simplified force fields. The value

of refining the calculated conformers using restrained MD (rMD) simulations with

state-of-art force fields is documented. This refinement however presents various

subtleties, from the proper formatting of conformational restraints to the

definition of suitable protocols.

RESULTS: We describe a web interface to set up and run calculations with the

AMBER package, which we called AMPS-NMR (AMBER-based Portal Server for NMR

structures). The interface allows the refinement of NMR structures through rMD.

Some predefined protocols are provided for this purpose, which can be

personalized; it is also possible to create an entirely new protocol. AMPS-NMR

can handle various restraint types. Standard rMD refinement in explicit water of

the structures of three different proteins are shown as examples. AMPS-NMR

additionally includes a workspace for the user to store different calculations.

As an ancillary service, a web interface to AnteChamber is available, enabling

the calculation of force field parameters for organic molecules such as ligands

in protein-ligand adducts.

AVAILABILITY AND IMPLEMENTATION: AMPS-NMR is embedded within the NMR services of

the WeNMR project and is available at

http://py-enmr.cerm.unifi.it/access/index/amps-nmr; its use requires registration

with a digital certificate.

CONTACT: ivanobertini@cerm.unifi.it

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr415

PMID: 21757462 [Indexed for MEDLINE]

1605. Bioinformatics. 2011 Sep 1;27(17):2441-2. doi: 10.1093/bioinformatics/btr400.

Epub 2011 Jul 4.

ModeRNA server: an online tool for modeling RNA 3D structures.

Rother M(1), Milanowska K, Puton T, Jeleniewicz J, Rother K, Bujnicki JM.

Author information:

(1)Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614

Poznan, Poland.

SUMMARY: The diverse functional roles of non-coding RNA molecules are determined

by their underlying structure. ModeRNA server is an online tool for RNA 3D

structure modeling by the comparative approach, based on a template RNA structure

and a user-defined target-template sequence alignment. It offers an option to

search for potential templates, given the target sequence. The server also

provides tools for analyzing, editing and formatting of RNA structure files. It

facilitates the use of the ModeRNA software and offers new options in comparison

to the standalone program.

AVAILABILITY AND IMPLEMENTATION: ModeRNA server was implemented using the Python

language and the Django web framework. It is freely available at

http://iimcb.genesilico.pl/modernaserver.

CONTACT: iamb@genesilico.pl.

DOI: 10.1093/bioinformatics/btr400

PMID: 21727140 [Indexed for MEDLINE]

1606. Bioinformatics. 2011 Sep 1;27(17):2422-5. doi: 10.1093/bioinformatics/btr389.

Epub 2011 Jun 27.

BRISK--research-oriented storage kit for biology-related data.

Tan A(1), Tripp B, Daley D.

Author information:

(1)James Hogg iCAPTURE Center, Department of Medicine, University of British

Columbia, Vancouver, BC, Canada V6Z1Y6.

MOTIVATION: In genetic science, large-scale international research collaborations

represent a growing trend. These collaborations have demanding and challenging

database, storage, retrieval and communication needs. These studies typically

involve demographic and clinical data, in addition to the results from numerous

genomic studies (omics studies) such as gene expression, eQTL, genome-wide

association and methylation studies, which present numerous challenges, thus the

need for data integration platforms that can handle these complex data

structures. Inefficient methods of data transfer and access control still plague

research collaboration. As science becomes more and more collaborative in nature,

the need for a system that adequately manages data sharing becomes paramount.

RESULTS: Biology-Related Information Storage Kit (BRISK) is a package of several

web-based data management tools that provide a cohesive data integration and

management platform. It was specifically designed to provide the architecture

necessary to promote collaboration and expedite data sharing between scientists.

AVAILABILITY AND IMPLEMENTATION: The software, documentation, Java source code

and demo are available at http://genapha.icapture.ubc.ca/brisk/index.jsp. BRISK

was developed in Java, and tested on an Apache Tomcat 6 server with a MySQL

database.

CONTACT: denise.daley@hli.ubc.ca.

DOI: 10.1093/bioinformatics/btr389

PMID: 21712248 [Indexed for MEDLINE]

1607. J Biosci. 2011 Sep;36(4):709-17.

Eu-Detect: an algorithm for detecting eukaryotic sequences in metagenomic data

sets.

Mohammed MH(1), Chadaram S, Komanduri D, Ghosh TS, Mande SS.

Author information:

(1)Bio-Sciences R and D Division, TCS Innovation Labs, Tata Consultancy Services

Limited, Hyderabad 500081, India.

Physical partitioning techniques are routinely employed (during sample

preparation stage) for segregating the prokaryotic and eukaryotic fractions of

metagenomic samples. In spite of these efforts, several metagenomic studies

focusing on bacterial and archaeal populations have reported the presence of

contaminating eukaryotic sequences in metagenomic data sets. Contaminating

sequences originate not only from genomes of micro-eukaryotic species but also

from genomes of (higher) eukaryotic host cells. The latter scenario usually

occurs in the case of host-associated metagenomes. Identification and removal of

contaminating sequences is important, since these sequences not only impact

estimates of microbial diversity but also affect the accuracy of several

downstream analyses. Currently, the computational techniques used for identifying

contaminating eukaryotic sequences, being alignment based, are slow, inefficient,

and require huge computing resources. In this article, we present Eu-Detect, an

alignment-free algorithm that can rapidly identify eukaryotic sequences

contaminating metagenomic data sets. Validation results indicate that on a

desktop with modest hardware specifications, the Eu-Detect algorithm is able to

rapidly segregate DNA sequence fragments of prokaryotic and eukaryotic origin,

with high sensitivity. A Web server for the Eu-Detect algorithm is available at

http://metagenomics.atc.tcs.com/Eu-Detect/.

PMID: 21857117 [Indexed for MEDLINE]

1608. J Mol Graph Model. 2011 Sep;30:129-34. doi: 10.1016/j.jmgm.2011.06.014. Epub 2011

Jul 7.

OligoPred: a web-server for predicting homo-oligomeric proteins by incorporating

discrete wavelet transform into Chou's pseudo amino acid composition.

Qiu JD(1), Suo SB, Sun XY, Shi SP, Liang RP.

Author information:

(1)Department of Chemistry, Nanchang University, Nanchang 330031, China.

jdqiu@ncu.edu.cn

In vivo, some proteins exist as monomers (single polypeptide chains) and others

as oligomers. Not like monomers, oligomers are composed of two or more chains

(subunits) that are associated with each other through non-covalent interactions

and, occasionally, through disulfide bonds. These proteins are the structural

components of various biological functions, including cooperative effects,

allosteric mechanisms and ion-channel gating. However, with the dramatic increase

in the number of protein sequences submitted to the public data bank, it is

important for both basic research and drug discovery research to acquire the

possible knowledge about homo-oligomeric attributes of their interested proteins

in a timely manner. In this paper, a high-throughput method, combined support

vector machines with discrete wavelet transform, has been developed to predict

the protein homo-oligomers. The total accuracy obtained by the re-substitution

test, jackknife test and independent dataset test are 99.94%, 96.17% and 96.18%,

respectively, showing that the proposed method of extracting feature from the

protein sequences is effective and feasible for predicting homo-oligomers. The

online service is available at http://bioinfo.ncu.edu.cn/Services.aspx.

Copyright © 2011 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.jmgm.2011.06.014

PMID: 21802968 [Indexed for MEDLINE]

1609. J Struct Biol. 2011 Sep;175(3):264-76. doi: 10.1016/j.jsb.2011.04.016. Epub 2011

May 4.

BCL::EM-Fit: rigid body fitting of atomic structures into density maps using

geometric hashing and real space refinement.

Woetzel N(1), Lindert S, Stewart PL, Meiler J.

Author information:

(1)Department of Chemistry, Vanderbilt University, TN 37212, USA.

Cryo-electron microscopy (cryoEM) can visualize large macromolecular assemblies

at resolutions often below 10Å and recently as good as 3.8-4.5 Å. These density

maps provide important insights into the biological functioning of molecular

machineries such as viruses or the ribosome, in particular if atomic-resolution

crystal structures or models of individual components of the assembly can be

placed into the density map. The present work introduces a novel algorithm termed

BCL::EM-Fit that accurately fits atomic-detail structural models into medium

resolution density maps. In an initial step, a "geometric hashing" algorithm

provides a short list of likely placements. In a follow up Monte Carlo/Metropolis

refinement step, the initial placements are optimized by their cross correlation

coefficient. The resolution of density maps for a reliable fit was determined to

be 10 Å or better using tests with simulated density maps. The algorithm was

applied to fitting of capsid proteins into an experimental cryoEM density map of

human adenovirus at a resolution of 6.8 and 9.0 Å, and fitting of the GroEL

protein at 5.4 Å. In the process, the handedness of the cryoEM density map was

unambiguously identified. The BCL::EM-Fit algorithm offers an alternative to the

established Fourier/Real space fitting programs. BCL::EM-Fit is free for academic

use and available from a web server or as downloadable binary file at

http://www.meilerlab.org.

Copyright © 2011 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.jsb.2011.04.016

PMCID: PMC3150432

PMID: 21565271 [Indexed for MEDLINE]

1610. RNA Biol. 2011 Sep-Oct;8(5):922-34. doi: 10.4161/rna.8.5.16026. Epub 2011 Sep 1.

mirExplorer: detecting microRNAs from genome and next generation sequencing data

using the AdaBoost method with transition probability matrix and combined

features.

Guan DG(1), Liao JY, Qu ZH, Zhang Y, Qu LH.

Author information:

(1)State Key Laboratory of Biocontrol, Key Laboratory of Gene Engineering of the

Ministry of Education, Sun Yat-Sen University, Guangzhou, China.

microRNAs (miRNAs) represent an abundant group of small regulatory non-coding

RNAs in eukaryotes. The emergence of Next-generation sequencing (NGS)

technologies has allowed the systematic detection of small RNAs (sRNAs) and de

novo sequencing of genomes quickly and with low cost. As a result, there is an

increased need to develop fast miRNA prediction tools to annotate miRNAs from

various organisms with a high level of accuracy, using the genome sequence or the

NGS data. Several miRNA predictors have been proposed to achieve this purpose.

However, the accuracy and fitness for multiple species of existing predictors

needed to be improved. Here, we present a novel prediction tool called

mirExplorer, which is based on an integrated adaptive boosting method and

contains two modules. The first module named mirExplorer-genome was designed to

de novo predict pre-miRNAs from genome, and the second module named

mirExplorer-NGS was used to discover miRNAs from NGS data. A set of novel

features of pre-miRNA secondary structure and miRNA biogenesis has been extracted

to distinguish real pre-miRNAs from pseudo ones. We used outer-ten-fold

cross-validation to verify the mirExplorer-genome computation, which obtained a

specificity of 95.03% and a sensitivity of 93.71% on human data. This computation

was made on test data from 16 species, and it achieved an overall accuracy of

95.53%. Systematic outer-ten-fold cross-validation of the mirExplorer-NGS model

achieved a specificity of 98.3% and a sensitivity of 97.72%. We found that the

good performance of the mirExplorer-NGS model was upheld across species from

vertebrates to plants in test datasets. The mirExplorer is available as both web

server and software package at http://biocenter.sysu.edu.cn/mir/.

DOI: 10.4161/rna.8.5.16026

PMID: 21881406 [Indexed for MEDLINE]

1611. Bioinformatics. 2011 Aug 15;27(16):2302-3. doi: 10.1093/bioinformatics/btr385.

Epub 2011 Jun 23.

ODORactor: a web server for deciphering olfactory coding.

Liu X(1), Su X, Wang F, Huang Z, Wang Q, Li Z, Zhang R, Wu L, Pan Y, Chen Y,

Zhuang H, Chen G, Shi T, Zhang J.

Author information:

(1)Department of Pathophysiology, Key Laboratory of Cell Differentiation and

Apoptosis of National Ministry of Education, Rui-Jin Hospital, Shanghai Jiao-Tong

University School of Medicine, Shanghai, 200025, China.

SUMMARY: ODORactor is an open access web server aimed at providing a platform for

identifying odorant receptors (ORs) for small molecules and for browsing existing

OR-ligand pairs. It enables the prediction of ORs from the molecular structures

of arbitrary chemicals by integrating two individual functionalities: odorant

verification and OR recognition. The prediction of the ORs for several odorants

was experimentally validated in the study. In addition, ODORactor features a

comprehensive repertoire of olfactory information that has been manually curated

from literature. Therefore, ODORactor may provide an effective way to decipher

olfactory coding and could be a useful server tool for both basic olfaction

research in academia and for odorant discovery in industry.

AVAILABILITY: Freely available at http://mdl.shsmu.edu.cn/ODORactor

CONTACT: jian.zhang@sjtu.edu.cn

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr385

PMID: 21700676 [Indexed for MEDLINE]

1612. Bioinformatics. 2011 Aug 15;27(16):2209-15. doi: 10.1093/bioinformatics/btr374.

Epub 2011 Jun 23.

Gaia: automated quality assessment of protein structure models.

Kota P(1), Ding F, Ramachandran S, Dokholyan NV.

Author information:

(1)Department of Biochemistry and Biophysics, University of North Carolina at

Chapel Hill, NC 27599-7260, USA.

MOTIVATION: Increasing use of structural modeling for understanding

structure-function relationships in proteins has led to the need to ensure that

the protein models being used are of acceptable quality. Quality of a given

protein structure can be assessed by comparing various intrinsic structural

properties of the protein to those observed in high-resolution protein

structures.

RESULTS: In this study, we present tools to compare a given structure to

high-resolution crystal structures. We assess packing by calculating the total

void volume, the percentage of unsatisfied hydrogen bonds, the number of steric

clashes and the scaling of the accessible surface area. We assess covalent

geometry by determining bond lengths, angles, dihedrals and rotamers. The

statistical parameters for the above measures, obtained from high-resolution

crystal structures enable us to provide a quality-score that points to specific

areas where a given protein structural model needs improvement.

AVAILABILITY AND IMPLEMENTATION: We provide these tools that appraise protein

structures in the form of a web server Gaia (http://chiron.dokhlab.org). Gaia

evaluates the packing and covalent geometry of a given protein structure and

provides quantitative comparison of the given structure to high-resolution

crystal structures.

CONTACT: dokh@unc.edu

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr374

PMCID: PMC3150034

PMID: 21700672 [Indexed for MEDLINE]

1613. BMC Bioinformatics. 2011 Aug 3;12:317. doi: 10.1186/1471-2105-12-317.

NoD: a Nucleolar localization sequence detector for eukaryotic and viral

proteins.

Scott MS(1), Troshin PV, Barton GJ.

Author information:

(1)Division of Biological Chemistry and Drug Discovery, College of Life Sciences,

University of Dundee, Dow Street, Dundee DD1 5EH, UK.

michelle@compbio.dundee.ac.uk

BACKGROUND: Nucleolar localization sequences (NoLSs) are short targeting

sequences responsible for the localization of proteins to the nucleolus. Given

the large number of proteins experimentally detected in the nucleolus and the

central role of this subnuclear compartment in the cell, NoLSs are likely to be

important regulatory elements controlling cellular traffic. Although many

proteins have been reported to contain NoLSs, the systematic characterization of

this group of targeting motifs has only recently been carried out.

RESULTS: Here, we describe NoD, a web server and a command line program that

predicts the presence of NoLSs in proteins. Using the web server, users can

submit protein sequences through the NoD input form and are provided with a

graphical output of the NoLS score as a function of protein position. While the

web server is most convenient for making prediction for just a few proteins, the

command line version of NoD can return predictions for complete proteomes. NoD is

based on our recently described human-trained artificial neural network

predictor. Through stringent independent testing of the predictor using available

experimentally validated NoLS-containing eukaryotic and viral proteins, the NoD

sensitivity and positive predictive value were estimated to be 71% and 79%

respectively.

CONCLUSIONS: NoD is the first tool to provide predictions of nucleolar

localization sequences in diverse eukaryotes and viruses. NoD can be run

interactively online at http://www.compbio.dundee.ac.uk/nod or downloaded to use

locally.

DOI: 10.1186/1471-2105-12-317

PMCID: PMC3166288

PMID: 21812952 [Indexed for MEDLINE]

1614. BMC Bioinformatics. 2011 Aug 2;12:316. doi: 10.1186/1471-2105-12-316.

GSV: a web-based genome synteny viewer for customized data.

Revanna KV(1), Chiu CC, Bierschank E, Dong Q.

Author information:

(1)Department of Biological Sciences, University of North Texas, Denton, Texas

76203, USA.

BACKGROUND: The analysis of genome synteny is a common practice in comparative

genomics. With the advent of DNA sequencing technologies, individual biologists

can rapidly produce their genomic sequences of interest. Although web-based

synteny visualization tools are convenient for biologists to use, none of the

existing ones allow biologists to upload their own data for analysis.

RESULTS: We have developed the web-based Genome Synteny Viewer (GSV) that allows

users to upload two data files for synteny visualization, the mandatory synteny

file for specifying genomic positions of conserved regions and the optional

genome annotation file. GSV presents two selected genomes in a single integrated

view while still retaining the browsing flexibility necessary for exploring

individual genomes. Users can browse and filter for genomic regions of interest,

change the color or shape of each annotation track as well as re-order, hide or

show the tracks dynamically. Additional features include downloadable images,

immediate email notification and tracking of usage history. The entire GSV

package is also light-weighted which enables easy local installation.

CONCLUSIONS: GSV provides a unique option for biologists to analyze genome

synteny by uploading their own data set to a web-based comparative genome

browser. A web server hosting GSV is provided at

http://cas-bioinfo.cas.unt.edu/gsv, and the software is also freely available for

local installations.

DOI: 10.1186/1471-2105-12-316

PMCID: PMC3199762

PMID: 21810250 [Indexed for MEDLINE]

1615. Bioinformatics. 2011 Aug 1;27(15):2147-8. doi: 10.1093/bioinformatics/btr357.

Epub 2011 Jun 17.

CHASM and SNVBox: toolkit for detecting biologically important single nucleotide

mutations in cancer.

Wong WC(1), Kim D, Carter H, Diekhans M, Ryan MC, Karchin R.

Author information:

(1)Department of Biomedical Engineering and Institute for Computational Medicine,

Johns Hopkins University, Baltimore, MD, USA.

SUMMARY: Thousands of cancer exomes are currently being sequenced, yielding

millions of non-synonymous single nucleotide variants (SNVs) of possible

relevance to disease etiology. Here, we provide a software toolkit to prioritize

SNVs based on their predicted contribution to tumorigenesis. It includes a

database of precomputed, predictive features covering all positions in the

annotated human exome and can be used either stand-alone or as part of a larger

variant discovery pipeline.

AVAILABILITY AND IMPLEMENTATION: MySQL database, source code and binaries freely

available for academic/government use at http://wiki.chasmsoftware.org, Source in

Python and C++. Requires 32 or 64-bit Linux system (tested on Fedora Core 8,10,11

and Ubuntu 10), 2.5\*≤ Python <3.0\*, MySQL server >5.0, 60 GB available hard disk

space (50 MB for software and data files, 40 GB for MySQL database dump when

uncompressed), 2 GB of RAM.

DOI: 10.1093/bioinformatics/btr357

PMCID: PMC3137226

PMID: 21685053 [Indexed for MEDLINE]

1616. Bioinformatics. 2011 Aug 1;27(15):2076-82. doi: 10.1093/bioinformatics/btr350.

Epub 2011 Jun 11.

Improving protein fold recognition and template-based modeling by employing

probabilistic-based matching between predicted one-dimensional structural

properties of query and corresponding native properties of templates.

Yang Y(1), Faraggi E, Zhao H, Zhou Y.

Author information:

(1)School of Informatics, Indiana University Purdue University, Indianapolis, IN

46202, USA.

MOTIVATION: In recent years, development of a single-method fold-recognition

server lags behind consensus and multiple template techniques. However, a good

consensus prediction relies on the accuracy of individual methods. This article

reports our efforts to further improve a single-method fold recognition technique

called SPARKS by changing the alignment scoring function and incorporating the

SPINE-X techniques that make improved prediction of secondary structure, backbone

torsion angle and solvent accessible surface area.

RESULTS: The new method called SPARKS-X was tested with the SALIGN benchmark for

alignment accuracy, Lindahl and SCOP benchmarks for fold recognition, and CASP 9

blind test for structure prediction. The method is compared to several

state-of-the-art techniques such as HHPRED and BoostThreader. Results show that

SPARKS-X is one of the best single-method fold recognition techniques. We further

note that incorporating multiple templates and refinement in model building will

likely further improve SPARKS-X.

AVAILABILITY: The method is available as a SPARKS-X server at

http://sparks.informatics.iupui.edu/

DOI: 10.1093/bioinformatics/btr350

PMCID: PMC3137224

PMID: 21666270 [Indexed for MEDLINE]

1617. Bioinformatics. 2011 Aug 1;27(15):2163-4. doi: 10.1093/bioinformatics/btr348.

Epub 2011 Jun 11.

Xwalk: computing and visualizing distances in cross-linking experiments.

Kahraman A(1), Malmström L, Aebersold R.

Author information:

(1)Department of Biology, Institute of Molecular Systems Biology, Swiss Federal

Institute of Technology (ETH Zurich), CH-8093 Zurich, Switzerland.

abdullah@imsb.biol.ethz.ch

MOTIVATION: Chemical cross-linking of proteins or protein complexes and the mass

spectrometry-based localization of the cross-linked amino acids in peptide

sequences is a powerful method for generating distance restraints on the

substrate's topology.

RESULTS: Here, we introduce the algorithm Xwalk for predicting and validating

these cross-links on existing protein structures. Xwalk calculates and displays

non-linear distances between chemically cross-linked amino acids on protein

surfaces, while mimicking the flexibility and non-linearity of cross-linker

molecules. It returns a 'solvent accessible surface distance', which corresponds

to the length of the shortest path between two amino acids, where the path leads

through solvent occupied space without penetrating the protein surface.

AVAILABILITY: Xwalk is freely available as a web server or stand-alone JAVA

application at http://www.xwalk.org.

DOI: 10.1093/bioinformatics/btr348

PMCID: PMC3137222

PMID: 21666267 [Indexed for MEDLINE]

1618. Bioinformatics. 2011 Aug 1;27(15):2151-2. doi: 10.1093/bioinformatics/btr338.

Epub 2011 Jun 8.

RNASAlign: RNA structural alignment system.

Wong TK(1), Wan KL, Hsu BY, Cheung BW, Hon WK, Lam TW, Yiu SM.

Author information:

(1)Department of Computer Science, The University of Hong Kong, Pokfulam Road,

Hong Kong.

MOTIVATION: Structural alignment of RNA is found to be a useful computational

technique for idenitfying non-coding RNAs (ncRNAs). However, existing tools do

not handle structures with pseudoknots. Although algorithms exist that can handle

structural alignment for different types of pseudoknots, no software tools are

available and users have to determine the type of pseudoknots to select the

appropriate algoirthm to use which limits the usage of structural alignment in

identifying novel ncRNAs.

RESULTS: We implemented the first web server, RNASAlign, which can automatically

identify the pseudoknot type of a secondary structure and perform structural

alignment of a folded RNA with every region of a target DNA/RNA sequence. Regions

with high similarity scores and low e-values, together with the detailed

alignments will be reported to the user. Experiments on more than 350 ncRNA

families show that RNASAlign is effective.

AVAILABILITY: http://www.bio8.cs.hku.hk/RNASAlign.

DOI: 10.1093/bioinformatics/btr338

PMID: 21659321 [Indexed for MEDLINE]

1619. Bioinformatics. 2011 Aug 1;27(15):2161-2. doi: 10.1093/bioinformatics/btr343.

Epub 2011 Jun 8.

The LabelHash server and tools for substructure-based functional annotation.

Moll M(1), Bryant DH, Kavraki LE.

Author information:

(1)Department of Computer Science, Rice University, Houston, TX 77005, USA.

mmoll@rice.edu

SUMMARY: The LabelHash server and tools are designed for large-scale substructure

comparison. The main use is to predict the function of unknown proteins. Given a

set of (putative) functional residues, LabelHash finds all occurrences of

matching substructures in the entire Protein Data Bank, along with a statistical

significance estimate and known functional annotations for each match. The

results can be downloaded for further analysis in any molecular viewer. For

Chimera, there is a plugin to facilitate this process.

AVAILABILITY: The web site is free and open to all users with no login

requirements at http://labelhash.kavrakilab.org

DOI: 10.1093/bioinformatics/btr343

PMID: 21659320 [Indexed for MEDLINE]

1620. Nucleic Acids Res. 2011 Aug;39(14):e93. doi: 10.1093/nar/gkr240. Epub 2011 May

18.

TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots.

Bon M(1), Orland H.

Author information:

(1)Institut de Physique Théorique, CEA Saclay, CNRS URA 2306, 91191

Gif-sur-Yvette, France.

We present TT2NE, a new algorithm to predict RNA secondary structures with

pseudoknots. The method is based on a classification of RNA structures according

to their topological genus. TT2NE is guaranteed to find the minimum free energy

structure regardless of pseudoknot topology. This unique proficiency is obtained

at the expense of the maximum length of sequences that can be treated, but

comparison with state-of-the-art algorithms shows that TT2NE significantly

improves the quality of predictions. Analysis of TT2NE's incorrect predictions

sheds light on the need to study how sterical constraints limit the range of

pseudoknotted structures that can be formed from a given sequence. An

implementation of TT2NE on a public server can be found at

http://ipht.cea.fr/rna/tt2ne.php.

DOI: 10.1093/nar/gkr240

PMCID: PMC3152363

PMID: 21593129 [Indexed for MEDLINE]

1621. J Clin Bioinforma. 2011 Jul 28;1(1):20. doi: 10.1186/2043-9113-1-20.

FISH Oracle: a web server for flexible visualization of DNA copy number data in a

genomic context.

Mader M(1), Simon R, Steinbiss S, Kurtz S.

Author information:

(1)Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146

Hamburg, Germany. kurtz@zbh.uni-hamburg.de.

BACKGROUND: The rapidly growing amount of array CGH data requires improved

visualization software supporting the process of identifying candidate cancer

genes. Optimally, such software should work across multiple microarray platforms,

should be able to cope with data from different sources and should be easy to

operate.

RESULTS: We have developed a web-based software FISH Oracle to visualize data

from multiple array CGH experiments in a genomic context. Its fast visualization

engine and advanced web and database technology supports highly interactive use.

FISH Oracle comes with a convenient data import mechanism, powerful search

options for genomic elements (e.g. gene names or karyobands), quick navigation

and zooming into interesting regions, and mechanisms to export the visualization

into different high quality formats. These features make the software especially

suitable for the needs of life scientists.

CONCLUSIONS: FISH Oracle offers a fast and easy to use visualization tool for

array CGH and SNP array data. It allows for the identification of genomic regions

representing minimal common changes based on data from one or more experiments.

FISH Oracle will be instrumental to identify candidate onco and tumor suppressor

genes based on the frequency and genomic position of DNA copy number changes. The

FISH Oracle application and an installed demo web server are available at

http://www.zbh.uni-hamburg.de/fishoracle.

DOI: 10.1186/2043-9113-1-20

PMCID: PMC3164613

PMID: 21884636

1622. BMC Bioinformatics. 2011 Jul 26;12:304. doi: 10.1186/1471-2105-12-304.

Applications of the pipeline environment for visual informatics and genomics

computations.

Dinov ID(1), Torri F, Macciardi F, Petrosyan P, Liu Z, Zamanyan A, Eggert P,

Pierce J, Genco A, Knowles JA, Clark AP, Van Horn JD, Ames J, Kesselman C, Toga

AW.

Author information:

(1)Laboratory of Neuro Imaging (LONI), University of California, Los Angeles, Los

Angeles, CA 90095, USA.

BACKGROUND: Contemporary informatics and genomics research require efficient,

flexible and robust management of large heterogeneous data, advanced

computational tools, powerful visualization, reliable hardware infrastructure,

interoperability of computational resources, and detailed data and

analysis-protocol provenance. The Pipeline is a client-server distributed

computational environment that facilitates the visual graphical construction,

execution, monitoring, validation and dissemination of advanced data analysis

protocols.

RESULTS: This paper reports on the applications of the LONI Pipeline environment

to address two informatics challenges - graphical management of diverse genomics

tools, and the interoperability of informatics software. Specifically, this

manuscript presents the concrete details of deploying general informatics suites

and individual software tools to new hardware infrastructures, the design,

validation and execution of new visual analysis protocols via the Pipeline

graphical interface, and integration of diverse informatics tools via the

Pipeline eXtensible Markup Language syntax. We demonstrate each of these

processes using several established informatics packages (e.g., miBLAST, EMBOSS,

mrFAST, GWASS, MAQ, SAMtools, Bowtie) for basic local sequence alignment and

search, molecular biology data analysis, and genome-wide association studies.

These examples demonstrate the power of the Pipeline graphical workflow

environment to enable integration of bioinformatics resources which provide a

well-defined syntax for dynamic specification of the input/output parameters and

the run-time execution controls.

CONCLUSIONS: The LONI Pipeline environment http://pipeline.loni.ucla.edu provides

a flexible graphical infrastructure for efficient biomedical computing and

distributed informatics research. The interactive Pipeline resource manager

enables the utilization and interoperability of diverse types of informatics

resources. The Pipeline client-server model provides computational power to a

broad spectrum of informatics investigators--experienced developers and novice

users, user with or without access to advanced computational-resources (e.g.,

Grid, data), as well as basic and translational scientists. The open development,

validation and dissemination of computational networks (pipeline workflows)

facilitates the sharing of knowledge, tools, protocols and best practices, and

enables the unbiased validation and replication of scientific findings by the

entire community.

DOI: 10.1186/1471-2105-12-304

PMCID: PMC3199760

PMID: 21791102 [Indexed for MEDLINE]

1623. BMC Res Notes. 2011 Jul 20;4:237. doi: 10.1186/1756-0500-4-237.

Analysis and prediction of cancerlectins using evolutionary and domain

information.

Kumar R(1), Panwar B, Chauhan JS, Raghava GP.

Author information:

(1)Bioinformatics Centre Institute of Microbial Technology, Sector-39A,

Chandigarh, India. raghava@imtech.res.in.

BACKGROUND: Predicting the function of a protein is one of the major challenges

in the post-genomic era where a large number of protein sequences of unknown

function are accumulating rapidly. Lectins are the proteins that specifically

recognize and bind to carbohydrate moieties present on either proteins or lipids.

Cancerlectins are those lectins that play various important roles in tumor cell

differentiation and metastasis. Although the two types of proteins are linked,

still there is no computational method available that can distinguish

cancerlectins from the large pool of non-cancerlectins. Hence, it is imperative

to develop a method that can distinguish between cancer and non-cancerlectins.

RESULTS: All the models developed in this study are based on a non-redundant

dataset containing 178 cancerlectins and 226 non-cancerlectins in which no two

sequences have more than 50% sequence similarity. We have applied the similarity

search based technique, i.e. BLAST, and achieved a maximum accuracy of 43.25%.

The amino acids compositional analysis have shown that certain residues (e.g.

Leucine, Proline) were preferred in cancerlectins whereas some other (e.g.

Asparatic acid, Asparagine) were preferred in non-cancerlectins. It has been

found that the PROSITE domain "Crystalline beta gamma" was abundant in

cancerlectins whereas domains like "SUEL-type lectin domain" were found mainly in

non-cancerlectins. An SVM-based model has been developed to differentiate between

the cancer and non-cancerlectins which achieved a maximum Matthew's correlation

coefficient (MCC) value of 0.32 with an accuracy of 64.84%, using amino acid

compositions. We have developed a model based on dipeptide compositions which

achieved an MCC value of 0.30 with an accuracy of 64.84%. Thereafter, we have

developed models based on split compositions (2 and 4 parts) and achieved an MCC

value of 0.31, 0.32 with accuracies of 65.10% and 66.09%, respectively. An SVM

model based on Position Specific Scoring Matrix (PSSM), generated by PSI-BLAST,

was developed and achieved an MCC value of 0.36 with an accuracy of 68.34%.

Finally, we have integrated the PROSITE domain information with PSSM and

developed an SVM model that has achieved an MCC value of 0.38 with 69.09%

accuracy.

CONCLUSION: BLAST has been found inefficient to distinguish between cancer and

non-cancerlectins. We analyzed the protein sequences of cancer and

non-cancerlectins and identified interesting patterns. We have been able to

identify PROSITE domains that are preferred in cancer and non-cancerlectins and

thus provided interesting insights into the two types of proteins. The method

developed in this study will be useful for researchers studying cancerlectins,

lectins and cancer biology. The web-server based on the above study, is available

at http://www.imtech.res.in/raghava/cancer\_pred/

DOI: 10.1186/1756-0500-4-237

PMCID: PMC3161874

PMID: 21774797

1624. Bioinformatics. 2011 Jul 15;27(14):2018-20. doi: 10.1093/bioinformatics/btr333.

Epub 2011 Jun 2.

The FAF-Drugs2 server: a multistep engine to prepare electronic chemical compound

collections.

Lagorce D(1), Maupetit J, Baell J, Sperandio O, Tufféry P, Miteva MA, Galons H,

Villoutreix BO.

Author information:

(1)MTi, Ressource Parisienne en Bioinformatique Structurale, Institut National de

Santé et de Recherche Médicale (INSERM), UMR-S 973 - Paris Diderot University,

75205 Paris, Cedex 13, France. david.lagorce@inserm.fr

SUMMARY: The FAF-Drugs2 server is a web application that prepares chemical

compound libraries prior to virtual screening or that assists hit selection/lead

optimization before chemical synthesis or ordering. The FAF-Drugs2 web server is

an enhanced version of the FAF-Drugs2 package that now includes Pan Assay

Interference Compounds detection. This online toolkit has been designed through a

user-centered approach with emphasis on user-friendliness. This is a unique

online tool allowing to prepare large compound libraries with in house or

user-defined filtering parameters.

AVAILABILITY: The FAF-Drugs2 server is freely available at

http://bioserv.rpbs.univ-paris-diderot.fr/FAF-Drugs/.

DOI: 10.1093/bioinformatics/btr333

PMID: 21636592 [Indexed for MEDLINE]

1625. Bioinformatics. 2011 Jul 15;27(14):2001-2. doi: 10.1093/bioinformatics/btr304.

Epub 2011 May 18.

Java bioinformatics analysis web services for multiple sequence

alignment--JABAWS:MSA.

Troshin PV(1), Procter JB, Barton GJ.

Author information:

(1)Biological Chemistry and Drug Discovery, College of Life Sciences, University

of Dundee, Dundee DD1 5EH, UK.

SUMMARY: JABAWS is a web services framework that simplifies the deployment of web

services for bioinformatics. JABAWS:MSA provides services for five multiple

sequence alignment (MSA) methods (Probcons, T-coffee, Muscle, Mafft and

ClustalW), and is the system employed by the Jalview multiple sequence analysis

workbench since version 2.6. A fully functional, easy to set up server is

provided as a Virtual Appliance (VA), which can be run on most operating systems

that support a virtualization environment such as VMware or Oracle VirtualBox.

JABAWS is also distributed as a Web Application aRchive (WAR) and can be

configured to run on a single computer and/or a cluster managed by Grid Engine,

LSF or other queuing systems that support DRMAA. JABAWS:MSA provides clients full

access to each application's parameters, allows administrators to specify named

parameter preset combinations and execution limits for each application through

simple configuration files. The JABAWS command-line client allows integration of

JABAWS services into conventional scripts.

AVAILABILITY AND IMPLEMENTATION: JABAWS is made freely available under the Apache

2 license and can be obtained from: http://www.compbio.dundee.ac.uk/jabaws.

DOI: 10.1093/bioinformatics/btr304

PMCID: PMC3129525

PMID: 21593132 [Indexed for MEDLINE]

1626. Bioinformatics. 2011 Jul 15;27(14):2003-5. doi: 10.1093/bioinformatics/btr191.

Epub 2011 Apr 14.

RING: networking interacting residues, evolutionary information and energetics in

protein structures.

Martin AJ(1), Vidotto M, Boscariol F, Di Domenico T, Walsh I, Tosatto SC.

Author information:

(1)Department of Biology, University of Padova, Viale G. Colombo 3, 35131 Padova,

Italy.

MOTIVATION: Residue interaction networks (RINs) have been used in the literature

to describe the protein 3D structure as a graph where nodes represent residues

and edges physico-chemical interactions, e.g. hydrogen bonds or van-der-Waals

contacts. Topological network parameters can be calculated over RINs and have

been correlated with various aspects of protein structure and function. Here we

present a novel web server, RING, to construct physico-chemically valid RINs

interactively from PDB files for subsequent visualization in the Cytoscape

platform. The additional structure-based parameters secondary structure, solvent

accessibility and experimental uncertainty can be combined with information

regarding residue conservation, mutual information and residue-based energy

scoring functions. Different visualization styles are provided to facilitate

visualization and standard plugins can be used to calculate topological

parameters in Cytoscape. A sample use case analyzing the active site of

glutathione peroxidase is presented.

AVAILABILITY: The RING server, supplementary methods, examples and tutorials are

available for non-commercial use at URL: http://protein.bio.unipd.it/ring/.

DOI: 10.1093/bioinformatics/btr191

PMID: 21493660 [Indexed for MEDLINE]

1627. J Clin Bioinforma. 2011 Jul 15;1(1):18. doi: 10.1186/2043-9113-1-18.

Automated generation of massive image knowledge collections using Microsoft Live

Labs Pivot to promote neuroimaging and translational research.

Viangteeravat T(1), Anyanwu MN, Ra Nagisetty V, Kuscu E.

Author information:

(1)Clinical and Translational Science Institute University of Tennessee Health

Science Center, Memphis, TN 38163, USA. manyanwu@uthsc.edu.

BACKGROUND: Massive datasets comprising high-resolution images, generated in

neuro-imaging studies and in clinical imaging research, are increasingly

challenging our ability to analyze, share, and filter such images in clinical and

basic translational research. Pivot collection exploratory analysis provides each

user the ability to fully interact with the massive amounts of visual data to

fully facilitate sufficient sorting, flexibility and speed to fluidly access,

explore or analyze the massive image data sets of high-resolution images and

their associated meta information, such as neuro-imaging databases from the Allen

Brain Atlas. It is used in clustering, filtering, data sharing and classifying of

the visual data into various deep zoom levels and meta information categories to

detect the underlying hidden pattern within the data set that has been used.

METHOD: We deployed prototype Pivot collections using the Linux CentOS running on

the Apache web server. We also tested the prototype Pivot collections on other

operating systems like Windows (the most common variants) and UNIX, etc. It is

demonstrated that the approach yields very good results when compared with other

approaches used by some researchers for generation, creation, and clustering of

massive image collections such as the coronal and horizontal sections of the

mouse brain from the Allen Brain Atlas.

RESULTS: Pivot visual analytics was used to analyze a prototype of dataset Dab2

co-expressed genes from the Allen Brain Atlas. The metadata along with

high-resolution images were automatically extracted using the Allen Brain Atlas

API. It is then used to identify the hidden information based on the various

categories and conditions applied by using options generated from automated

collection. A metadata category like chromosome, as well as data for individual

cases like sex, age, and plan attributes of a particular gene, is used to filter,

sort and to determine if there exist other genes with a similar characteristics

to Dab2. And online access to the mouse brain pivot collection can be viewed

using the link

http://edtech-dev.uthsc.edu/CTSI/teeDev1/unittest/PaPa/collection.html (user

name: tviangte and password: demome)

CONCLUSIONS: Our proposed algorithm has automated the creation of large image

Pivot collections; this will enable investigators of clinical research projects

to easily and quickly analyse the image collections through a perspective that is

useful for making critical decisions about the image patterns discovered.

DOI: 10.1186/2043-9113-1-18

PMCID: PMC3164611

PMID: 21884637

1628. BMC Pharmacol. 2011 Jul 6;11:5. doi: 10.1186/1471-2210-11-5.

A web server for predicting inhibitors against bacterial target GlmU protein.

Singla D(1), Anurag M, Dash D, Raghava GP.

Author information:

(1)Institute of Microbial Technology, Chandigarh, India. raghava@imtech.res.in

BACKGROUND: The emergence of drug resistant tuberculosis poses a serious concern

globally and researchers are in rigorous search for new drugs to fight against

these dreadful bacteria. Recently, the bacterial GlmU protein, involved in

peptidoglycan, lipopolysaccharide and techoic acid synthesis, has been identified

as an important drug target. A unique C-terminal disordered tail, essential for

survival and the absence of gene in host makes GlmU a suitable target for

inhibitor design.

RESULTS: This study describes the models developed for predicting inhibitory

activity (IC50) of chemical compounds against GlmU protein using QSAR and docking

techniques. These models were trained on 84 diverse compounds (GlmU inhibitors)

taken from PubChem BioAssay (AID 1376). These inhibitors were docked in the

active site of the C-terminal domain of GlmU protein (2OI6) using the AutoDock. A

QSAR model was developed using docking energies as descriptors and achieved

maximum correlation of 0.35/0.12 (r/r2) between actual and predicted pIC50.

Secondly, QSAR models were developed using molecular descriptors calculated using

various software packages and achieved maximum correlation of 0.77/0.60 (r/r2).

Finally, hybrid models were developed using various types of descriptors and

achieved high correlation of 0.83/0.70 (r/r2) between predicted and actual pIC50.

It was observed that some molecular descriptors used in this study had high

correlation with pIC50. We screened chemical libraries using models developed in

this study and predicted 40 potential GlmU inhibitors. These inhibitors could be

used to develop drugs against Mycobacterium tuberculosis.

CONCLUSION: These results demonstrate that docking energies can be used as

descriptors for developing QSAR models. The current work suggests that docking

energies based descriptors could be used along with commonly used molecular

descriptors for predicting inhibitory activity (IC50) of molecules against GlmU.

Based on this study an open source platform, http://crdd.osdd.net/raghava/gdoq,

has been developed for predicting inhibitors GlmU.

DOI: 10.1186/1471-2210-11-5

PMCID: PMC3146400

PMID: 21733180 [Indexed for MEDLINE]

1629. BMC Bioinformatics. 2011 Jul 5;12:275. doi: 10.1186/1471-2105-12-275.

AlignHUSH: alignment of HMMs using structure and hydrophobicity information.

Krishnadev O(1), Srinivasan N.

Author information:

(1)Molecular Biophysics Unit Indian Institute of Science, Bangalore 560012,

India.

BACKGROUND: Sensitive remote homology detection and accurate alignments

especially in the midnight zone of sequence similarity are needed for better

function annotation and structural modeling of proteins. An algorithm, AlignHUSH

for HMM-HMM alignment has been developed which is capable of recognizing

distantly related domain families The method uses structural information, in the

form of predicted secondary structure probabilities, and hydrophobicity of amino

acids to align HMMs of two sets of aligned sequences. The effect of using

adjoining column(s) information has also been investigated and is found to

increase the sensitivity of HMM-HMM alignments and remote homology detection.

RESULTS: We have assessed the performance of AlignHUSH using known evolutionary

relationships available in SCOP. AlignHUSH performs better than the best HMM-HMM

alignment methods and is observed to be even more sensitive at higher error

rates. Accuracy of the alignments obtained using AlignHUSH has been assessed

using the structure-based alignments available in BaliBASE. The alignment length

and the alignment quality are found to be appropriate for homology modeling and

function annotation. The alignment accuracy is found to be comparable to existing

methods for profile-profile alignments.

CONCLUSIONS: A new method to align HMMs has been developed and is shown to have

better sensitivity at error rates of 10% and above when compared to other

available programs. The proposed method could effectively aid obtaining clues to

functions of proteins of yet unknown function. A web-server incorporating the

AlignHUSH method is available at http://crick.mbu.iisc.ernet.in/~alignhush/

DOI: 10.1186/1471-2105-12-275

PMCID: PMC3228556

PMID: 21729312 [Indexed for MEDLINE]

1630. Bioinformatics. 2011 Jul 1;27(13):i85-93. doi: 10.1093/bioinformatics/btr215.

IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots

using integer programming.

Sato K(1), Kato Y, Hamada M, Akutsu T, Asai K.

Author information:

(1)Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha,

Kashiwa, Chiba 277-8561, Japan. satoken@k.u-tokyo.ac.jp

MOTIVATION: Pseudoknots found in secondary structures of a number of functional

RNAs play various roles in biological processes. Recent methods for predicting

RNA secondary structures cover certain classes of pseudoknotted structures, but

only a few of them achieve satisfying predictions in terms of both speed and

accuracy.

RESULTS: We propose IPknot, a novel computational method for predicting RNA

secondary structures with pseudoknots based on maximizing expected accuracy of a

predicted structure. IPknot decomposes a pseudoknotted structure into a set of

pseudoknot-free substructures and approximates a base-pairing probability

distribution that considers pseudoknots, leading to the capability of modeling a

wide class of pseudoknots and running quite fast. In addition, we propose a

heuristic algorithm for refining base-paring probabilities to improve the

prediction accuracy of IPknot. The problem of maximizing expected accuracy is

solved by using integer programming with threshold cut. We also extend IPknot so

that it can predict the consensus secondary structure with pseudoknots when a

multiple sequence alignment is given. IPknot is validated through extensive

experiments on various datasets, showing that IPknot achieves better prediction

accuracy and faster running time as compared with several competitive prediction

methods.

AVAILABILITY: The program of IPknot is available at

http://www.ncrna.org/software/ipknot/. IPknot is also available as a web server

at http://rna.naist.jp/ipknot/.

CONTACT: satoken@k.u-tokyo.ac.jp; ykato@is.naist.jp

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btr215

PMCID: PMC3117384

PMID: 21685106 [Indexed for MEDLINE]

1631. Bioinformatics. 2011 Jul 1;27(13):1878-9. doi: 10.1093/bioinformatics/btr278.

Epub 2011 May 5.

MPEA--metabolite pathway enrichment analysis.

Kankainen M(1), Gopalacharyulu P, Holm L, Oresic M.

Author information:

(1)VTT Technical Research Centre of Finland, Espooi, Finland.

matti.kankainen@helsinki.fi

We present metabolite pathway enrichment analysis (MPEA) for the visualization

and biological interpretation of metabolite data at the system level. Our tool

follows the concept of gene set enrichment analysis (GSEA) and tests whether

metabolites involved in some predefined pathway occur towards the top (or bottom)

of a ranked query compound list. In particular, MPEA is designed to handle

many-to-many relationships that may occur between the query compounds and

metabolite annotations. For a demonstration, we analysed metabolite profiles of

14 twin pairs with differing body weights. MPEA found significant pathways from

data that had no significant individual query compounds, its results were

congruent with those discovered from transcriptomics data and it detected more

pathways than the competing metabolic pathway method did.AVAILABILITY: The web

server and source code of MPEA are available at

http://ekhidna.biocenter.helsinki.fi/poxo/mpea/.

DOI: 10.1093/bioinformatics/btr278

PMID: 21551139 [Indexed for MEDLINE]

1632. IEEE/ACM Trans Comput Biol Bioinform. 2011 Jul-Aug;8(4):1067-79. doi:

10.1109/TCBB.2010.94.

Predicting MHC-II binding affinity using multiple instance regression.

EL-Manzalawy Y(1), Dobbs D, Honavar V.

Author information:

(1)Department of Systems and Computers Engineering, Al-Azhar University, Cairo,

Egypt. yasser@azhar.edu.eg

Reliably predicting the ability of antigen peptides to bind to major

histocompatibility complex class II (MHC-II) molecules is an essential step in

developing new vaccines. Uncovering the amino acid sequence correlates of the

binding affinity of MHC-II binding peptides is important for understanding

pathogenesis and immune response. The task of predicting MHC-II binding peptides

is complicated by the significant variability in their length. Most existing

computational methods for predicting MHC-II binding peptides focus on identifying

a nine amino acids core region in each binding peptide. We formulate the problems

of qualitatively and quantitatively predicting flexible length MHC-II peptides as

multiple instance learning and multiple instance regression problems,

respectively. Based on this formulation, we introduce MHCMIR, a novel method for

predicting MHC-II binding affinity using multiple instance regression. We present

results of experiments using several benchmark data sets that show that MHCMIR is

competitive with the state-of-the-art methods for predicting MHC-II binding

peptides. An online web server that implements the MHCMIR method for MHC-II

binding affinity prediction is freely accessible at

http://ailab.cs.iastate.edu/mhcmir.

DOI: 10.1109/TCBB.2010.94

PMCID: PMC3400677

PMID: 20855923 [Indexed for MEDLINE]

1633. J Bacteriol. 2011 Jul;193(13):3228-40. doi: 10.1128/JB.00350-11. Epub 2011 Apr

29.

Inference of the transcriptional regulatory network in Staphylococcus aureus by

integration of experimental and genomics-based evidence.

Ravcheev DA(1), Best AA, Tintle N, Dejongh M, Osterman AL, Novichkov PS, Rodionov

DA.

Author information:

(1)Sanford-Burnham Medical Research Institute, La Jolla, CA 92037, USA.

Transcriptional regulatory networks are fine-tuned systems that help

microorganisms respond to changes in the environment and cell physiological

state. We applied the comparative genomics approach implemented in the RegPredict

Web server combined with SEED subsystem analysis and available information on

known regulatory interactions for regulatory network reconstruction for the human

pathogen Staphylococcus aureus and six related species from the family

Staphylococcaceae. The resulting reference set of 46 transcription factor

regulons contains more than 1,900 binding sites and 2,800 target genes involved

in the central metabolism of carbohydrates, amino acids, and fatty acids;

respiration; the stress response; metal homeostasis; drug and metal resistance;

and virulence. The inferred regulatory network in S. aureus includes ∼320

regulatory interactions between 46 transcription factors and ∼550 candidate

target genes comprising 20% of its genome. We predicted ∼170 novel interactions

and 24 novel regulons for the control of the central metabolic pathways in S.

aureus. The reconstructed regulons are largely variable in the Staphylococcaceae:

only 20% of S. aureus regulatory interactions are conserved across all studied

genomes. We used a large-scale gene expression data set for S. aureus to assess

relationships between the inferred regulons and gene expression patterns. The

predicted reference set of regulons is captured within the Staphylococcus

collection in the RegPrecise database (http://regprecise.lbl.gov).

DOI: 10.1128/JB.00350-11

PMCID: PMC3133287

PMID: 21531804 [Indexed for MEDLINE]

1634. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W118-24. doi:

10.1093/nar/gkr432.

ncFANs: a web server for functional annotation of long non-coding RNAs.

Liao Q(1), Xiao H, Bu D, Xie C, Miao R, Luo H, Zhao G, Yu K, Zhao H, Skogerbø G,

Chen R, Wu Z, Liu C, Zhao Y.

Author information:

(1)Bioinformatics Research Group, Key Laboratory of Intelligent Information

Processing, Institute of Computing Technology, Chinese Academy of Sciences,

Beijing, PR China.

Recent interest in the non-coding transcriptome has resulted in the

identification of large numbers of long non-coding RNAs (lncRNAs) in mammalian

genomes, most of which have not been functionally characterized. Computational

exploration of the potential functions of these lncRNAs will therefore facilitate

further work in this field of research. We have developed a practical and

user-friendly web interface called ncFANs (non-coding RNA Function ANnotation

server), which is the first web service for functional annotation of human and

mouse lncRNAs. On the basis of the re-annotated Affymetrix microarray data,

ncFANs provides two alternative strategies for lncRNA functional annotation: one

utilizing three aspects of a coding-non-coding gene co-expression (CNC) network,

the other identifying condition-related differentially expressed lncRNAs. ncFANs

introduces a highly efficient way of re-using the abundant pre-existing

microarray data. The present version of ncFANs includes re-annotated CDF files

for 10 human and mouse Affymetrix microarrays, and the server will be

continuously updated with more re-annotated microarray platforms and lncRNA data.

ncFANs is freely accessible at http://www.ebiomed.org/ncFANs/ or

http://www.noncode.org/ncFANs/.

DOI: 10.1093/nar/gkr432

PMCID: PMC3125796

PMID: 21715382 [Indexed for MEDLINE]

1635. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W316-22. doi:

10.1093/nar/gkr483.

KOBAS 2.0: a web server for annotation and identification of enriched pathways

and diseases.

Xie C(1), Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L.

Author information:

(1)Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene

Research, College of Life Sciences, Peking University, Beijing, China.

High-throughput experimental technologies often identify dozens to hundreds of

genes related to, or changed in, a biological or pathological process. From these

genes one wants to identify biological pathways that may be involved and diseases

that may be implicated. Here, we report a web server, KOBAS 2.0, which annotates

an input set of genes with putative pathways and disease relationships based on

mapping to genes with known annotations. It allows for both ID mapping and

cross-species sequence similarity mapping. It then performs statistical tests to

identify statistically significantly enriched pathways and diseases. KOBAS 2.0

incorporates knowledge across 1327 species from 5 pathway databases (KEGG

PATHWAY, PID, BioCyc, Reactome and Panther) and 5 human disease databases (OMIM,

KEGG DISEASE, FunDO, GAD and NHGRI GWAS Catalog). KOBAS 2.0 can be accessed at

http://kobas.cbi.pku.edu.cn.

DOI: 10.1093/nar/gkr483

PMCID: PMC3125809

PMID: 21715386 [Indexed for MEDLINE]

1636. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W38-44. doi: 10.1093/nar/gkr441.

FFAS server: novel features and applications.

Jaroszewski L(1), Li Z, Cai XH, Weber C, Godzik A.

Author information:

(1)Bioinformatics and Systems Biology Program, Sanford Burnham Medical Research

Institute, 10901 N. Torrey Pines Road, La Jolla, CA 92037, USA.

The Fold and Function Assignment System (FFAS) server [Jaroszewski et al. (2005)

FFAS03: a server for profile-profile sequence alignments. Nucleic Acids Research,

33, W284-W288] implements the algorithm for protein profile-profile alignment

introduced originally in [Rychlewski et al. (2000) Comparison of sequence

profiles. Strategies for structural predictions using sequence information.

Protein Science: a Publication of the Protein Society, 9, 232-241]. Here, we

present updates, changes and novel functionality added to the server since 2005

and discuss its new applications. The sequence database used to calculate

sequence profiles was enriched by adding sets of publicly available metagenomic

sequences. The profile of a user's protein can now be compared with ∼20

additional profile databases, including several complete proteomes, human

proteins involved in genetic diseases and a database of microbial virulence

factors. A newly developed interface uses a system of tabs, allowing the user to

navigate multiple results pages, and also includes novel functionality, such as a

dotplot graph viewer, modeling tools, an improved 3D alignment viewer and links

to the database of structural similarities. The FFAS server was also optimized

for speed: running times were reduced by an order of magnitude. The FFAS server,

http://ffas.godziklab.org, has no log-in requirement, albeit there is an option

to register and store results in individual, password-protected directories.

Source code and Linux executables for the FFAS program are available for download

from the FFAS server.

DOI: 10.1093/nar/gkr441

PMCID: PMC3125803

PMID: 21715387 [Indexed for MEDLINE]

1637. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W3-7. doi: 10.1093/nar/gkr514.

The 2011 Bioinformatics Links Directory update: more resources, tools and

databases and features to empower the bioinformatics community.

Brazas MD(1), Yim DS, Yamada JT, Ouellette BF.

Author information:

(1)Ontario Institute for Cancer Research, 101 College St., Suite 800, Toronto,

Ontario, Canada M5G 0A3.

The Bioinformatics Links Directory continues its collaboration with Nucleic Acids

Research to collaboratively publish and compile a freely accessible, online

collection of tools, databases and resource materials for bioinformatics and

molecular biology research. The July 2011 Web Server issue of Nucleic Acids

Research adds an additional 78 web server tools and 14 updates to the directory

at http://bioinformatics.ca/links\_directory/.

DOI: 10.1093/nar/gkr514

PMCID: PMC3125814

PMID: 21715385 [Indexed for MEDLINE]

1638. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W167-70. doi:

10.1093/nar/gkr490.

MultiFit: a web server for fitting multiple protein structures into their

electron microscopy density map.

Tjioe E(1), Lasker K, Webb B, Wolfson HJ, Sali A.

Author information:

(1)Department of Bioengineering and Therapeutic Sciences, University of

California, San Francisco, CA, USA.

Advances in electron microscopy (EM) allow for structure determination of large

biological assemblies at increasingly higher resolutions. A key step in this

process is fitting multiple component structures into an EM-derived density map

of their assembly. Here, we describe a web server for this task. The server takes

as input a set of protein structures in the PDB format and an EM density map in

the MRC format. The output is an ensemble of models ranked by their quality of

fit to the density map. The models can be viewed online or downloaded from the

website. The service is available at; http://salilab.org/multifit/ and

http://bioinfo3d.cs.tau.ac.il/.

DOI: 10.1093/nar/gkr490

PMCID: PMC3125811

PMID: 21715383 [Indexed for MEDLINE]

1639. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W177-83. doi:

10.1093/nar/gkr482. Epub 2011 Jun 21.

KINARI-Web: a server for protein rigidity analysis.

Fox N(1), Jagodzinski F, Li Y, Streinu I.

Author information:

(1)Department of Computer Science, 140 Governors Drive, University of

Massachusetts, Amherst, MA 01003, USA.

KINARI-Web is an interactive web server for performing rigidity analysis and

visually exploring rigidity properties of proteins. It also provides tools for

pre-processing the input data, such as selecting relevant chains from PDB files,

adding hydrogen atoms and identifying stabilizing interactions. KINARI-Web offers

a quick-start option for beginners, and highly customizable features for the

experienced user. Chains, residues or atoms, as well as stabilizing constraints

can be selected, removed or added, and the user can designate how different

chemical interactions should be modeled during rigidity analysis. The enhanced

Jmol-based visualizer allows for zooming in, highlighting or investigating

different calculated rigidity properties of a molecular structure. KINARI-Web is

freely available at http://kinari.cs.umass.edu.

DOI: 10.1093/nar/gkr482

PMCID: PMC3125808

PMID: 21693559 [Indexed for MEDLINE]

1640. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W357-61. doi:

10.1093/nar/gkr468. Epub 2011 Jun 14.

MarkUs: a server to navigate sequence-structure-function space.

Fischer M(1), Zhang QC, Dey F, Chen BY, Honig B, Petrey D.

Author information:

(1)Howard Hughes Medical Institute, Department of Biochemistry and Molecular

Biophysics, Center for Computational Biology and Bioinformatics, Columbia

University, New York, NY 10032, USA.

We describe MarkUs, a web server for analysis and comparison of the structural

and functional properties of proteins. In contrast to a 'structure in/function

out' approach to protein function annotation, the server is designed to be highly

interactive and to allow flexibility in the examination of possible functions,

suggested either automatically by various similarity measures or specified by a

user directly. This is combined with tools that allow a user to assess

independently whether or not a suggested function is consistent with the

bioinformatic and biophysical properties of a given query structure, further

allowing the user to generate testable hypotheses. The server is available at

http://wiki.c2b2.columbia.edu/honiglab\_public/index.php/Software:Mark-Us.

DOI: 10.1093/nar/gkr468

PMCID: PMC3125806

PMID: 21672961 [Indexed for MEDLINE]

1641. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W149-54. doi:

10.1093/nar/gkr467. Epub 2011 Jun 14.

RNApredator: fast accessibility-based prediction of sRNA targets.

Eggenhofer F(1), Tafer H, Stadler PF, Hofacker IL.

Author information:

(1)Institute of Theoretical Chemistry, University of Vienna, Währingerstrasse 17,

A-1090 Vienna, Austria. egg@tbi.univie.ac.at

Bacterial genomes encode a plethora of small RNAs (sRNAs), which are

heterogeneous in size, structure and function. Most sRNAs act as

post-transcriptional regulators by means of specific base pairing interactions

with the 5'-untranslated region of mRNA transcripts, thereby modifying the

stability of the target transcript and/or its ability to be translated. Here, we

present RNApredator, a web server for the prediction of sRNA targets. The user

can choose from a set of over 2155 genomes and plasmids from 1183 bacterial

species. RNApredator then uses a dynamic programming approach, RNAplex, to

compute putative targets. Compared to web servers with a similar task,

RNApredator takes the accessibility of the target during the target search into

account, improving the specificity of the predictions. Furthermore, enrichment in

Gene Ontology terms, cellular pathways as well as changes in accessibilities

along the target sequence can be done in fully automated post-processing steps.

The predictive performance of the underlying dynamic programming approach RNAplex

is similar to that of more complex methods, but needs at least three orders of

magnitude less time to complete. RNApredator is available at

http://rna.tbi.univie.ac.at/RNApredator.

DOI: 10.1093/nar/gkr467

PMCID: PMC3125805

PMID: 21672960 [Indexed for MEDLINE]

1642. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W235-41. doi:

10.1093/nar/gkr437. Epub 2011 Jun 14.

firestar--advances in the prediction of functionally important residues.

Lopez G(1), Maietta P, Rodriguez JM, Valencia A, Tress ML.

Author information:

(1)Structural Computational Biology Group, Spanish National Cancer Research

Centre (CNIO), c. Melchor Fernandez Almagro, 3, 28029 Madrid, Spain.

firestar is a server for predicting catalytic and ligand-binding residues in

protein sequences. Here, we present the important developments since the first

release of firestar. Previous versions of the server required human

interpretation of the results; the server is now fully automatized. firestar has

been implemented as a web service and can now be run in high-throughput mode.

Prediction coverage has been greatly improved with the extension of the FireDB

database and the addition of alignments generated by HHsearch. Ligands in FireDB

are now classified for biological relevance. Many of the changes have been

motivated by the critical assessment of techniques for protein structure

prediction (CASP) ligand-binding prediction experiment, which provided us with a

framework to test the performance of firestar. URL:

http://firedb.bioinfo.cnio.es/Php/FireStar.php.

DOI: 10.1093/nar/gkr437

PMCID: PMC3125799

PMID: 21672959 [Indexed for MEDLINE]

1643. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W347-52. doi:

10.1093/nar/gkr485. Epub 2011 Jun 14.

PHAST: a fast phage search tool.

Zhou Y(1), Liang Y, Lynch KH, Dennis JJ, Wishart DS.

Author information:

(1)Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

T6G 2E8.

PHAge Search Tool (PHAST) is a web server designed to rapidly and accurately

identify, annotate and graphically display prophage sequences within bacterial

genomes or plasmids. It accepts either raw DNA sequence data or partially

annotated GenBank formatted data and rapidly performs a number of database

comparisons as well as phage 'cornerstone' feature identification steps to

locate, annotate and display prophage sequences and prophage features. Relative

to other prophage identification tools, PHAST is up to 40 times faster and up to

15% more sensitive. It is also able to process and annotate both raw DNA sequence

data and Genbank files, provide richly annotated tables on prophage features and

prophage 'quality' and distinguish between intact and incomplete prophage. PHAST

also generates downloadable, high quality, interactive graphics that display all

identified prophage components in both circular and linear genomic views. PHAST

is available at (http://phast.wishartlab.com).

DOI: 10.1093/nar/gkr485

PMCID: PMC3125810

PMID: 21672955 [Indexed for MEDLINE]

1644. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W254-60. doi:

10.1093/nar/gkr434. Epub 2011 Jun 11.

PAComplex: a web server to infer peptide antigen families and binding models from

TCR-pMHC complexes.

Liu IH(1), Lo YS, Yang JM.

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu, Taiwan.

One of the most adaptive immune responses is triggered by specific T-cell

receptors (TCR) binding to peptide-major histocompatibility complexes (pMHC).

Despite the availability of many prediction servers to identify peptides binding

to MHC, these servers are often lacking in peptide-TCR interactions and detailed

atomic interacting models. PAComplex is the first web server investigating both

pMHC and peptide-TCR interfaces to infer peptide antigens and homologous peptide

antigens of a query. This server first identifies significantly similar TCR-pMHC

templates (joint Z-value ≥ 4.0) of the query by using antibody-antigen and

protein-protein interacting scoring matrices for peptide-TCR and pMHC interfaces,

respectively. PAComplex then identifies the homologous peptide antigens of these

hit templates from complete pathogen genome databases (≥10(8) peptide candidates

from 864,628 protein sequences of 389 pathogens) and experimental peptide

databases (80,057 peptides in 2287 species). Finally, the server outputs peptide

antigens and homologous peptide antigens of the query and displays detailed

interacting models (e.g. hydrogen bonds and steric interactions in two

interfaces) of hitTCR-pMHC templates. Experimental results demonstrate that the

proposed server can achieve high prediction accuracy and offer potential peptide

antigens across pathogens. We believe that the server is able to provide valuable

insights for the peptide vaccine and MHC restriction. The PAComplex sever is

available at http://PAcomplex.life.nctu.edu.tw.

DOI: 10.1093/nar/gkr434

PMCID: PMC3125798

PMID: 21666259 [Indexed for MEDLINE]

1645. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W223-8. doi: 10.1093/nar/gkr412.

Epub 2011 Jun 11.

GPU.proton.DOCK: Genuine Protein Ultrafast proton equilibria consistent DOCKing.

Kantardjiev AA(1).

Author information:

(1)Biophysical Chemistry Group, Institute of Organic Chemistry, Bulgarian Academy

of Sciences, and Department of Physics, Sofia University, Sofia, Bulgaria.

alexkant@orgchm.bas.bg

GPU.proton.DOCK (Genuine Protein Ultrafast proton equilibria consistent DOCKing)

is a state of the art service for in silico prediction of protein-protein

interactions via rigorous and ultrafast docking code. It is unique in providing

stringent account of electrostatic interactions self-consistency and proton

equilibria mutual effects of docking partners. GPU.proton.DOCK is the first

server offering such a crucial supplement to protein docking algorithms--a step

toward more reliable and high accuracy docking results. The code (especially the

Fast Fourier Transform bottleneck and electrostatic fields computation) is

parallelized to run on a GPU supercomputer. The high performance will be of use

for large-scale structural bioinformatics and systems biology projects, thus

bridging physics of the interactions with analysis of molecular networks. We

propose workflows for exploring in silico charge mutagenesis effects. Special

emphasis is given to the interface-intuitive and user-friendly. The input is

comprised of the atomic coordinate files in PDB format. The advanced user is

provided with a special input section for addition of non-polypeptide charges,

extra ionogenic groups with intrinsic pK(a) values or fixed ions. The output is

comprised of docked complexes in PDB format as well as interactive visualization

in a molecular viewer. GPU.proton.DOCK server can be accessed at

http://gpudock.orgchm.bas.bg/.

DOI: 10.1093/nar/gkr412

PMCID: PMC3125792

PMID: 21666258 [Indexed for MEDLINE]

1646. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W184-9. doi: 10.1093/nar/gkr430.

Epub 2011 Jun 10.

AquaSAXS: a web server for computation and fitting of SAXS profiles with

non-uniformally hydrated atomic models.

Poitevin F(1), Orland H, Doniach S, Koehl P, Delarue M.

Author information:

(1)Institut Pasteur, Unit of Structural Dynamics of Macromolecules, CNRS, URA

2185, 91191 Gif/Yvette Cedex, France. fpoitevi@pasteur.fr

Small Angle X-ray Scattering (SAXS) techniques are becoming more and more useful

for structural biologists and biochemists, thanks to better access to dedicated

synchrotron beamlines, better detectors and the relative easiness of sample

preparation. The ability to compute the theoretical SAXS profile of a given

structural model, and to compare this profile with the measured scattering

intensity, yields crucial structural informations about the macromolecule under

study and/or its complexes in solution. An important contribution to the profile,

besides the macromolecule itself and its solvent-excluded volume, is the excess

density due to the hydration layer. AquaSAXS takes advantage of recently

developed methods, such as AquaSol, that give the equilibrium solvent density map

around macromolecules, to compute an accurate SAXS/WAXS profile of a given

structure and to compare it to the experimental one. Here, we describe the

interface architecture and capabilities of the AquaSAXS web server

(http://lorentz.dynstr.pasteur.fr/aquasaxs.php).

DOI: 10.1093/nar/gkr430

PMCID: PMC3125794

PMID: 21665925 [Indexed for MEDLINE]

1647. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W203-9. doi: 10.1093/nar/gkr410.

Epub 2011 Jun 10.

SA-Mot: a web server for the identification of motifs of interest extracted from

protein loops.

Regad L(1), Saladin A, Maupetit J, Geneix C, Camproux AC.

Author information:

(1)INSERM, U973, Université Paris 7-Paris Diderot, UMR-S973, MTi F-75013 Paris,

France. leslie.regad@univ-paris-diderot.fr

The detection of functional motifs is an important step for the determination of

protein functions. We present here a new web server SA-Mot (Structural Alphabet

Motif) for the extraction and location of structural motifs of interest from

protein loops. Contrary to other methods, SA-Mot does not focus only on

functional motifs, but it extracts recurrent and conserved structural motifs

involved in structural redundancy of loops. SA-Mot uses the structural word

notion to extract all structural motifs from uni-dimensional sequences

corresponding to loop structures. Then, SA-Mot provides a description of these

structural motifs using statistics computed in the loop data set and in SCOP

superfamily, sequence and structural parameters. SA-Mot results correspond to an

interactive table listing all structural motifs extracted from a target structure

and their associated descriptors. Using this information, the users can easily

locate loop regions that are important for the protein folding and function. The

SA-Mot web server is available at http://sa-mot.mti.univ-paris-diderot.fr.

DOI: 10.1093/nar/gkr410

PMCID: PMC3125790

PMID: 21665924 [Indexed for MEDLINE]

1648. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W499-504. doi:

10.1093/nar/gkr413. Epub 2011 Jun 10.

Detecting selection in immunoglobulin sequences.

Uduman M(1), Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH.

Author information:

(1)Interdepartmental Program in Computational Biology and Bioinformatics, Yale

University, New Haven, CT 06520, USA.

The ability to detect selection by analyzing mutation patterns in experimentally

derived immunoglobulin (Ig) sequences is a critical part of many studies. Such

techniques are useful not only for understanding the response to pathogens, but

also to determine the role of antigen-driven selection in autoimmunity, B cell

cancers and the diversification of pre-immune repertoires in certain species.

Despite its importance, quantifying selection in experimentally derived sequences

is fraught with difficulties. The necessary parameters for statistical tests

(such as the expected frequency of replacement mutations in the absence of

selection) are non-trivial to calculate, and results are not easily interpretable

when analyzing more than a handful of sequences. We have developed a web server

that implements our previously proposed Focused binomial test for detecting

selection. Several features are integrated into the web site in order to

facilitate analysis, including V(D)J germline segment identification with IMGT

alignment, batch submission of sequences and integration of additional test

statistics proposed by other groups. We also implement a Z-score-based statistic

that increases the power of detecting selection while maintaining specificity,

and further allows for the combined analysis of sequences from different

germlines. The tool is freely available at http://clip.med.yale.edu/selection.

DOI: 10.1093/nar/gkr413

PMCID: PMC3125793

PMID: 21665923 [Indexed for MEDLINE]

1649. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W368-74. doi:

10.1093/nar/gkr440. Epub 2011 Jun 7.

PILGRM: an interactive data-driven discovery platform for expert biologists.

Greene CS(1), Troyanskaya OG.

Author information:

(1)Department of Computer Science, Lewis-Sigler Institute for Integrative

Genomics, Princeton University, Princeton, NJ 08544, USA. csgreene@princeton.edu

PILGRM (the platform for interactive learning by genomics results mining) puts

advanced supervised analysis techniques applied to enormous gene expression

compendia into the hands of bench biologists. This flexible system empowers its

users to answer diverse biological questions that are often outside of the scope

of common databases in a data-driven manner. This capability allows domain

experts to quickly and easily generate hypotheses about biological processes,

tissues or diseases of interest. Specifically PILGRM helps biologists generate

these hypotheses by analyzing the expression levels of known relevant genes in

large compendia of microarray data. Because PILGRM is data-driven, it complements

a user's knowledge and literature analysis with mining of diverse functional

genomic data, thereby generating novel predictions that can drive experimental

follow-up. This server is free, does not require registration and is available

for use at http://pilgrm.princeton.edu.

DOI: 10.1093/nar/gkr440

PMCID: PMC3125802

PMID: 21653547 [Indexed for MEDLINE]

1650. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W307-15. doi:

10.1093/nar/gkr378. Epub 2011 Jun 6.

g:Profiler--a web server for functional interpretation of gene lists (2011

update).

Reimand J(1), Arak T, Vilo J.

Author information:

(1)University of Tartu, Institute of Computer Science, Tartu, Estonia.

juri.reimand@ut.ee

Functional interpretation of candidate gene lists is an essential task in modern

biomedical research. Here, we present the 2011 update of g:Profiler

(http://biit.cs.ut.ee/gprofiler/), a popular collection of web tools for

functional analysis. g:GOSt and g:Cocoa combine comprehensive methods for

interpreting gene lists, ordered lists and list collections in the context of

biomedical ontologies, pathways, transcription factor and microRNA regulatory

motifs and protein-protein interactions. Additional tools, namely the biomolecule

ID mapping service (g:Convert), gene expression similarity searcher (g:Sorter)

and gene homology searcher (g:Orth) provide numerous ways for further analysis

and interpretation. In this update, we have implemented several features of

interest to the community: (i) functional analysis of single nucleotide

polymorphisms and other DNA polymorphisms is supported by chromosomal queries;

(ii) network analysis identifies enriched protein-protein interaction modules in

gene lists; (iii) functional analysis covers human disease genes; and (iv)

improved statistics and filtering provide more concise results. g:Profiler is a

regularly updated resource that is available for a wide range of species,

including mammals, plants, fungi and insects.

DOI: 10.1093/nar/gkr378

PMCID: PMC3125778

PMID: 21646343 [Indexed for MEDLINE]

1651. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W190-6. doi: 10.1093/nar/gkr411.

Epub 2011 Jun 6.

CSpritz: accurate prediction of protein disorder segments with annotation for

homology, secondary structure and linear motifs.

Walsh I(1), Martin AJ, Di Domenico T, Vullo A, Pollastri G, Tosatto SC.

Author information:

(1)Department of Biology, University of Padua, Padova 35131, Italy.

CSpritz is a web server for the prediction of intrinsic protein disorder. It is a

combination of previous Spritz with two novel orthogonal systems developed by our

group (Punch and ESpritz). Punch is based on sequence and structural templates

trained with support vector machines. ESpritz is an efficient single sequence

method based on bidirectional recursive neural networks. Spritz was extended to

filter predictions based on structural homologues. After extensive testing,

predictions are combined by averaging their probabilities. The CSpritz website

can elaborate single or multiple predictions for either short or long disorder.

The server provides a global output page, for download and simultaneous

statistics of all predictions. Links are provided to each individual protein

where the amino acid sequence and disorder prediction are displayed along with

statistics for the individual protein. As a novel feature, CSpritz provides

information about structural homologues as well as secondary structure and short

functional linear motifs in each disordered segment. Benchmarking was performed

on the very recent CASP9 data, where CSpritz would have ranked consistently well

with a Sw measure of 49.27 and AUC of 0.828. The server, together with help and

methods pages including examples, are freely available at URL:

http://protein.bio.unipd.it/cspritz/.

DOI: 10.1093/nar/gkr411

PMCID: PMC3125791

PMID: 21646342 [Indexed for MEDLINE]

1652. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W270-7. doi: 10.1093/nar/gkr366.

Epub 2011 May 29.

SwissDock, a protein-small molecule docking web service based on EADock DSS.

Grosdidier A(1), Zoete V, Michielin O.

Author information:

(1)Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Génopode, CH-1015

Lausanne, Switzerland.

Most life science processes involve, at the atomic scale, recognition between two

molecules. The prediction of such interactions at the molecular level, by

so-called docking software, is a non-trivial task. Docking programs have a wide

range of applications ranging from protein engineering to drug design. This

article presents SwissDock, a web server dedicated to the docking of small

molecules on target proteins. It is based on the EADock DSS engine, combined with

setup scripts for curating common problems and for preparing both the target

protein and the ligand input files. An efficient Ajax/HTML interface was designed

and implemented so that scientists can easily submit dockings and retrieve the

predicted complexes. For automated docking tasks, a programmatic SOAP interface

has been set up and template programs can be downloaded in Perl, Python and PHP.

The web site also provides an access to a database of manually curated complexes,

based on the Ligand Protein Database. A wiki and a forum are available to the

community to promote interactions between users. The SwissDock web site is

available online at http://www.swissdock.ch. We believe it constitutes a step

toward generalizing the use of docking tools beyond the traditional molecular

modeling community.

DOI: 10.1093/nar/gkr366

PMCID: PMC3125772

PMID: 21624888 [Indexed for MEDLINE]

1653. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W249-53. doi:

10.1093/nar/gkr431. Epub 2011 May 27.

Rosetta FlexPepDock web server--high resolution modeling of peptide-protein

interactions.

London N(1), Raveh B, Cohen E, Fathi G, Schueler-Furman O.

Author information:

(1)Department of Microbiology and Molecular Genetics, Institute for Medical

Research Israel-Canada, Hadassah Medical School, The Hebrew University, Jerusalem

91120, Israel.

Peptide-protein interactions are among the most prevalent and important

interactions in the cell, but a large fraction of those interactions lack

detailed structural characterization. The Rosetta FlexPepDock web server

(http://flexpepdock.furmanlab.cs.huji.ac.il/) provides an interface to a

high-resolution peptide docking (refinement) protocol for the modeling of

peptide-protein complexes, implemented within the Rosetta framework. Given a

protein receptor structure and an approximate, possibly inaccurate model of the

peptide within the receptor binding site, the FlexPepDock server refines the

peptide to high resolution, allowing full flexibility to the peptide backbone and

to all side chains. This protocol was extensively tested and benchmarked on a

wide array of non-redundant peptide-protein complexes, and was proven effective

when applied to peptide starting conformations within 5.5 Å backbone root mean

square deviation from the native conformation. FlexPepDock has been applied to

several systems that are mediated and regulated by peptide-protein interactions.

This easy to use and general web server interface allows non-expert users to

accurately model their specific peptide-protein interaction of interest.

DOI: 10.1093/nar/gkr431

PMCID: PMC3125795

PMID: 21622962 [Indexed for MEDLINE]

1654. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W155-9. doi: 10.1093/nar/gkr319.

Epub 2011 May 27.

psRNATarget: a plant small RNA target analysis server.

Dai X(1), Zhao PX.

Author information:

(1)Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble

Parkway, Ardmore, OK 73401, USA.

Plant endogenous non-coding short small RNAs (20-24 nt), including microRNAs

(miRNAs) and a subset of small interfering RNAs (ta-siRNAs), play important role

in gene expression regulatory networks (GRNs). For example, many transcription

factors and development-related genes have been reported as targets of these

regulatory small RNAs. Although a number of miRNA target prediction algorithms

and programs have been developed, most of them were designed for animal miRNAs

which are significantly different from plant miRNAs in the target recognition

process. These differences demand the development of separate plant miRNA (and

ta-siRNA) target analysis tool(s). We present psRNATarget, a plant small RNA

target analysis server, which features two important analysis functions: (i)

reverse complementary matching between small RNA and target transcript using a

proven scoring schema, and (ii) target-site accessibility evaluation by

calculating unpaired energy (UPE) required to 'open' secondary structure around

small RNA's target site on mRNA. The psRNATarget incorporates recent discoveries

in plant miRNA target recognition, e.g. it distinguishes translational and

post-transcriptional inhibition, and it reports the number of small RNA/target

site pairs that may affect small RNA binding activity to target transcript. The

psRNATarget server is designed for high-throughput analysis of next-generation

data with an efficient distributed computing back-end pipeline that runs on a

Linux cluster. The server front-end integrates three simplified user-friendly

interfaces to accept user-submitted or preloaded small RNAs and transcript

sequences; and outputs a comprehensive list of small RNA/target pairs along with

the online tools for batch downloading, key word searching and results sorting.

The psRNATarget server is freely available at

http://plantgrn.noble.org/psRNATarget/.

DOI: 10.1093/nar/gkr319

PMCID: PMC3125753

PMID: 21622958 [Indexed for MEDLINE]

1655. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W437-43. doi:

10.1093/nar/gkr391. Epub 2011 May 27.

ICSNPathway: identify candidate causal SNPs and pathways from genome-wide

association study by one analytical framework.

Zhang K(1), Chang S, Cui S, Guo L, Zhang L, Wang J.

Author information:

(1)Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of

Sciences, Beijing 100101, China.

Genome-wide association study (GWAS) is widely utilized to identify genes

involved in human complex disease or some other trait. One key challenge for GWAS

data interpretation is to identify causal SNPs and provide profound evidence on

how they affect the trait. Currently, researches are focusing on identification

of candidate causal variants from the most significant SNPs of GWAS, while there

is lack of support on biological mechanisms as represented by pathways. Although

pathway-based analysis (PBA) has been designed to identify disease-related

pathways by analyzing the full list of SNPs from GWAS, it does not emphasize on

interpreting causal SNPs. To our knowledge, so far there is no web server

available to solve the challenge for GWAS data interpretation within one

analytical framework. ICSNPathway is developed to identify candidate causal SNPs

and their corresponding candidate causal pathways from GWAS by integrating

linkage disequilibrium (LD) analysis, functional SNP annotation and PBA.

ICSNPathway provides a feasible solution to bridge the gap between GWAS and

disease mechanism study by generating hypothesis of SNP → gene → pathway(s). The

ICSNPathway server is freely available at http://icsnpathway.psych.ac.cn/.

DOI: 10.1093/nar/gkr391

PMCID: PMC3125783

PMID: 21622953 [Indexed for MEDLINE]

1656. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W197-202. doi:

10.1093/nar/gkr292. Epub 2011 May 26.

BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural

annotation of protein sequences.

Piovesan D(1), Martelli PL, Fariselli P, Zauli A, Rossi I, Casadio R.

Author information:

(1)Department of Biology, Bologna Biocomputing Group, Bologna Computational

Biology Network, Bologna, Italy.

We introduce BAR-PLUS (BAR(+)), a web server for functional and structural

annotation of protein sequences. BAR(+) is based on a large-scale genome cross

comparison and a non-hierarchical clustering procedure characterized by a metric

that ensures a reliable transfer of features within clusters. In this version,

the method takes advantage of a large-scale pairwise sequence comparison of

13,495,736 protein chains also including 988 complete proteomes. Available

sequence annotation is derived from UniProtKB, GO, Pfam and PDB. When PDB

templates are present within a cluster (with or without their SCOP

classification), profile Hidden Markov Models (HMMs) are computed on the basis of

sequence to structure alignment and are cluster-associated (Cluster-HMM).

Therefrom, a library of 10,858 HMMs is made available for aligning even distantly

related sequences for structural modelling. The server also provides pairwise

query sequence-structural target alignments computed from the correspondent

Cluster-HMM. BAR(+) in its present version allows three main categories of

annotation: PDB [with or without SCOP (\*)] and GO and/or Pfam; PDB (\*) without GO

and/or Pfam; GO and/or Pfam without PDB (\*) and no annotation. Each category can

further comprise clusters where GO and Pfam functional annotations are or are not

statistically significant. BAR(+) is available at

http://bar.biocomp.unibo.it/bar2.0.

DOI: 10.1093/nar/gkr292

PMCID: PMC3125743

PMID: 21622657 [Indexed for MEDLINE]

1657. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W518-23. doi:

10.1093/nar/gkr388. Epub 2011 May 26.

CoMet--a web server for comparative functional profiling of metagenomes.

Lingner T(1), Asshauer KP, Schreiber F, Meinicke P.

Author information:

(1)Department of Bioinformatics, Institute for Microbiology and Genetics,

Georg-August University Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany.

thomas@gobics.de

Analyzing the functional potential of newly sequenced genomes and metagenomes has

become a common task in biomedical and biological research. With the advent of

high-throughput sequencing technologies comparative metagenomics opens the way to

elucidate the genetically determined similarities and differences of complex

microbial communities. We developed the web server 'CoMet'

(http://comet.gobics.de), which provides an easy-to-use comparative metagenomics

platform that is well-suitable for the analysis of large collections of

metagenomic short read data. CoMet combines the ORF finding and subsequent

assignment of protein sequences to Pfam domain families with a comparative

statistical analysis. Besides comprehensive tabular data files, the CoMet server

also provides visually interpretable output in terms of hierarchical clustering

and multi-dimensional scaling plots and thus allows a quick overview of a given

set of metagenomic samples.

DOI: 10.1093/nar/gkr388

PMCID: PMC3125781

PMID: 21622656 [Indexed for MEDLINE]

1658. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W278-82. doi:

10.1093/nar/gkr389. Epub 2011 May 26.

Phosfinder: a web server for the identification of phosphate-binding sites on

protein structures.

Parca L(1), Mangone I, Gherardini PF, Ausiello G, Helmer-Citterich M.

Author information:

(1)Centre for Molecular Bioinformatics, Department of Biology, University of Rome

Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy.

Phosfinder is a web server for the identification of phosphate binding sites in

protein structures. Phosfinder uses a structural comparison algorithm to scan a

query structure against a set of known 3D phosphate binding motifs. Whenever a

structural similarity between the query protein and a phosphate binding motif is

detected, the phosphate bound by the known motif is added to the protein

structure thus representing a putative phosphate binding site. Predicted binding

sites are then evaluated according to (i) their position with respect to the

query protein solvent-excluded surface and (ii) the conservation of the binding

residues in the protein family. The server accepts as input either the PDB code

of the protein to be analyzed or a user-submitted structure in PDB format. All

the search parameters are user modifiable. Phosfinder outputs a list of predicted

binding sites with detailed information about their structural similarity with

known phosphate binding motifs, and the conservation of the residues involved. A

graphical applet allows the user to visualize the predicted binding sites on the

query protein structure. The results on a set of 52 apo/holo structure pairs show

that the performance of our method is largely unaffected by ligand-induced

conformational changes. Phosfinder is available at

http://phosfinder.bio.uniroma2.it.

DOI: 10.1093/nar/gkr389

PMCID: PMC3125782

PMID: 21622655 [Indexed for MEDLINE]

1659. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W56-60. doi: 10.1093/nar/gkr402.

Epub 2011 May 26.

SLiMSearch 2.0: biological context for short linear motifs in proteins.

Davey NE(1), Haslam NJ, Shields DC, Edwards RJ.

Author information:

(1)Structural and Computational Biology Unit, European Molecular Biology

Laboratory, 69117 Heidelberg, Germany.

Short, linear motifs (SLiMs) play a critical role in many biological processes.

The SLiMSearch 2.0 (Short, Linear Motif Search) web server allows researchers to

identify occurrences of a user-defined SLiM in a proteome, using conservation and

protein disorder context statistics to rank occurrences. User-friendly output and

visualizations of motif context allow the user to quickly gain insight into the

validity of a putatively functional motif occurrence. For each motif occurrence,

overlapping UniProt features and annotated SLiMs are displayed. Visualization

also includes annotated multiple sequence alignments surrounding each occurrence,

showing conservation and protein disorder statistics in addition to known and

predicted SLiMs, protein domains and known post-translational modifications. In

addition, enrichment of Gene Ontology terms and protein interaction partners are

provided as indicators of possible motif function. All web server results are

available for download. Users can search motifs against the human proteome or a

subset thereof defined by Uniprot accession numbers or GO term. The SLiMSearch

server is available at: http://bioware.ucd.ie/slimsearch2.html.

DOI: 10.1093/nar/gkr402

PMCID: PMC3125787

PMID: 21622654 [Indexed for MEDLINE]

1660. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W524-7. doi: 10.1093/nar/gkr373.

Epub 2011 May 24.

PRI-CAT: a web-tool for the analysis, storage and visualization of plant ChIP-seq

experiments.

Muiño JM(1), Hoogstraat M, van Ham RC, van Dijk AD.

Author information:

(1)Applied Bioinformatics, Plant Research International, PO Box 619, 6700 AP

Wageningen, The Netherlands. jose.muino@wur.nl

Although several tools for the analysis of ChIP-seq data have been published

recently, there is a growing demand, in particular in the plant research

community, for computational resources with which such data can be processed,

analyzed, stored, visualized and integrated within a single, user-friendly

environment. To accommodate this demand, we have developed PRI-CAT (Plant

Research International ChIP-seq analysis tool), a web-based workflow tool for the

management and analysis of ChIP-seq experiments. PRI-CAT is currently focused on

Arabidopsis, but will be extended with other plant species in the near future.

Users can directly submit their sequencing data to PRI-CAT for automated

analysis. A QuickLoad server compatible with genome browsers is implemented for

the storage and visualization of DNA-binding maps. Submitted datasets and results

can be made publicly available through PRI-CAT, a feature that will enable

community-based integrative analysis and visualization of ChIP-seq experiments.

Secondary analysis of data can be performed with the aid of GALAXY, an external

framework for tool and data integration. PRI-CAT is freely available at

http://www.ab.wur.nl/pricat. No login is required.

DOI: 10.1093/nar/gkr373

PMCID: PMC3125775

PMID: 21609962 [Indexed for MEDLINE]

1661. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W385-90. doi:

10.1093/nar/gkr284. Epub 2011 May 23.

Update of PROFEAT: a web server for computing structural and physicochemical

features of proteins and peptides from amino acid sequence.

Rao HB(1), Zhu F, Yang GB, Li ZR, Chen YZ.

Author information:

(1)College of Chemistry, Sichuan University, Chengdu, 610064, PR China.

Sequence-derived structural and physicochemical features have been extensively

used for analyzing and predicting structural, functional, expression and

interaction profiles of proteins and peptides. PROFEAT has been developed as a

web server for computing commonly used features of proteins and peptides from

amino acid sequence. To facilitate more extensive studies of protein and

peptides, numerous improvements and updates have been made to PROFEAT. We added

new functions for computing descriptors of protein-protein and protein-small

molecule interactions, segment descriptors for local properties of protein

sequences, topological descriptors for peptide sequences and small molecule

structures. We also added new feature groups for proteins and peptides

(pseudo-amino acid composition, amphiphilic pseudo-amino acid composition, total

amino acid properties and atomic-level topological descriptors) as well as for

small molecules (atomic-level topological descriptors). Overall, PROFEAT computes

11 feature groups of descriptors for proteins and peptides, and a feature group

of more than 400 descriptors for small molecules plus the derived features for

protein-protein and protein-small molecule interactions. Our computational

algorithms have been extensively tested and used in a number of published works

for predicting proteins of specific structural or functional classes,

protein-protein interactions, peptides of specific functions and quantitative

structure activity relationships of small molecules. PROFEAT is accessible free

of charge at http://bidd.cz3.nus.edu.sg/cgi-bin/prof/protein/profnew.cgi.

DOI: 10.1093/nar/gkr284

PMCID: PMC3125735

PMID: 21609959 [Indexed for MEDLINE]

1662. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W455-61. doi:

10.1093/nar/gkr246. Epub 2011 May 23.

Génie: literature-based gene prioritization at multi genomic scale.

Fontaine JF(1), Priller F, Barbosa-Silva A, Andrade-Navarro MA.

Author information:

(1)Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125

Berlin, Germany. jean-fred.fontaine@mdc-berlin.de

Biomedical literature is traditionally used as a way to inform scientists of the

relevance of genes in relation to a research topic. However many genes,

especially from poorly studied organisms, are not discussed in the literature.

Moreover, a manual and comprehensive summarization of the literature attached to

the genes of an organism is in general impossible due to the high number of genes

and abstracts involved. We introduce the novel Génie algorithm that overcomes

these problems by evaluating the literature attached to all genes in a genome and

to their orthologs according to a selected topic. Génie showed high precision (up

to 100%) and the best performance in comparison to other algorithms in most of

the benchmarks, especially when high sensitivity was required. Moreover, the

prioritization of zebrafish genes involved in heart development, using human and

mouse orthologs, showed high enrichment in differentially expressed genes from

microarray experiments. The Génie web server supports hundreds of species,

millions of genes and offers novel functionalities. Common run times below a

minute, even when analyzing the human genome with hundreds of thousands of

literature records, allows the use of Génie in routine lab work.AVAILABILITY:

http://cbdm.mdc-berlin.de/tools/genie/.

DOI: 10.1093/nar/gkr246

PMCID: PMC3125729

PMID: 21609954 [Indexed for MEDLINE]

1663. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W283-7. doi: 10.1093/nar/gkr311.

Epub 2011 May 23.

PredUs: a web server for predicting protein interfaces using structural

neighbors.

Zhang QC(1), Deng L, Fisher M, Guan J, Honig B, Petrey D.

Author information:

(1)Department of Biochemistry and Molecular Biophysics, Center for Computational

Biology and Bioinformatics, Howard Hughes Medical Institute, Columbia University,

1130 St. Nicholas Avenue, Room 815, New York, NY 10032, USA.

We describe PredUs, an interactive web server for the prediction of

protein-protein interfaces. Potential interfacial residues for a query protein

are identified by 'mapping' contacts from known interfaces of the query protein's

structural neighbors to surface residues of the query. We calculate a score for

each residue to be interfacial with a support vector machine. Results can be

visualized in a molecular viewer and a number of interactive features allow users

to tailor a prediction to a particular hypothesis. The PredUs server is available

at: http://wiki.c2b2.columbia.edu/honiglab\_public/index.php/Software:PredUs.

DOI: 10.1093/nar/gkr311

PMCID: PMC3125747

PMID: 21609948 [Indexed for MEDLINE]

1664. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W334-8. doi: 10.1093/nar/gkr289.

Epub 2011 May 20.

PINTA: a web server for network-based gene prioritization from expression data.

Nitsch D(1), Tranchevent LC, Gonçalves JP, Vogt JK, Madeira SC, Moreau Y.

Author information:

(1)Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit

Leuven, 3001 Leuven, Belgium.

PINTA (available at http://www.esat.kuleuven.be/pinta/; this web site is free and

open to all users and there is no login requirement) is a web resource for the

prioritization of candidate genes based on the differential expression of their

neighborhood in a genome-wide protein-protein interaction network. Our strategy

is meant for biological and medical researchers aiming at identifying novel

disease genes using disease specific expression data. PINTA supports both

candidate gene prioritization (starting from a user defined set of candidate

genes) as well as genome-wide gene prioritization and is available for five

species (human, mouse, rat, worm and yeast). As input data, PINTA only requires

disease specific expression data, whereas various platforms (e.g. Affymetrix) are

supported. As a result, PINTA computes a gene ranking and presents the results as

a table that can easily be browsed and downloaded by the user.

DOI: 10.1093/nar/gkr289

PMCID: PMC3125740

PMID: 21602267 [Indexed for MEDLINE]

1665. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W24-8. doi: 10.1093/nar/gkr393.

Epub 2011 May 20.

CLICK--topology-independent comparison of biomolecular 3D structures.

Nguyen MN(1), Tan KP, Madhusudhan MS.

Author information:

(1)Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore

138671.

Our server, CLICK: http://mspc.bii.a-star.edu.sg/click, is capable of

superimposing the 3D structures of any pair of biomolecules (proteins, DNA, RNA,

etc.). The server makes use of the Cartesian coordinates of the molecules with

the option of using other structural features such as secondary structure,

solvent accessible surface area and residue depth to guide the alignment. CLICK

first looks for cliques of points (3-7 residues) that are structurally similar in

the pair of structures to be aligned. Using these local similarities, a

one-to-one equivalence is charted between the residues of the two structures. A

least square fit then superimposes the two structures. Our method is especially

powerful in establishing protein relationships by detecting similarities in

structural subdomains, domains and topological variants. CLICK has been

extensively benchmarked and compared with other popular methods for protein and

RNA structural alignments. In most cases, CLICK alignments were statistically

significantly better in terms of structure overlap. The method also recognizes

conformational changes that may have occurred in structural domains or subdomains

in one structure with respect to the other. For this purpose, the server produces

complementary alignments to maximize the extent of detectable similarity. Various

examples showcase the utility of our web server.

DOI: 10.1093/nar/gkr393

PMCID: PMC3125785

PMID: 21602266 [Indexed for MEDLINE]

1666. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W125-31. doi:

10.1093/nar/gkr374. Epub 2011 May 20.

miRvestigator: web application to identify miRNAs responsible for co-regulated

gene expression patterns discovered through transcriptome profiling.

Plaisier CL(1), Bare JC, Baliga NS.

Author information:

(1)Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109-5234,

USA.

Transcriptome profiling studies have produced staggering numbers of gene

co-expression signatures for a variety of biological systems. A significant

fraction of these signatures will be partially or fully explained by

miRNA-mediated targeted transcript degradation. miRvestigator takes as input

lists of co-expressed genes from Caenorhabditis elegans, Drosophila melanogaster,

G. gallus, Homo sapiens, Mus musculus or Rattus norvegicus and identifies the

specific miRNAs that are likely to bind to 3' un-translated region (UTR)

sequences to mediate the observed co-regulation. The novelty of our approach is

the miRvestigator hidden Markov model (HMM) algorithm which systematically

computes a similarity P-value for each unique miRNA seed sequence from the miRNA

database miRBase to an overrepresented sequence motif identified within the

3'-UTR of the query genes. We have made this miRNA discovery tool accessible to

the community by integrating our HMM algorithm with a proven algorithm for de

novo discovery of miRNA seed sequences and wrapping these algorithms into a

user-friendly interface. Additionally, the miRvestigator web server also produces

a list of putative miRNA binding sites within 3'-UTRs of the query transcripts to

facilitate the design of validation experiments. The miRvestigator is freely

available at http://mirvestigator.systemsbiology.net.

DOI: 10.1093/nar/gkr374

PMCID: PMC3125776

PMID: 21602264 [Indexed for MEDLINE]

1667. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W139-44. doi:

10.1093/nar/gkr351. Epub 2011 May 19.

mirAct: a web tool for evaluating microRNA activity based on gene expression

data.

Liang Z(1), Zhou H, He Z, Zheng H, Wu J.

Author information:

(1)Hefei National Laboratory for Physical Sciences at Microscale and School of

Life Science, University of Science and Technology of China, 96 Jinzhai Road,

Hefei, Anhui 230027, China. liangzhi@ustc.edu.cn

MicroRNAs (miRNAs) are critical regulators in the complex cellular networks. The

mirAct web server (http://sysbio.ustc.edu.cn/software/mirAct) is a tool designed

to investigate miRNA activity based on gene-expression data by using the negative

regulation relationship between miRNAs and their target genes. mirAct supports

multiple-class data and enables clustering analysis based on computationally

determined miRNA activity. Here, we describe the framework of mirAct, demonstrate

its performance by comparing with other similar programs and exemplify its

applications using case studies.

DOI: 10.1093/nar/gkr351

PMCID: PMC3125759

PMID: 21596785 [Indexed for MEDLINE]

1668. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W160-6. doi: 10.1093/nar/gkr358.

Epub 2011 May 19.

corRna: a web server for predicting multiple-point deleterious mutations in

structural RNAs.

Lam E(1), Kam A, Waldispühl J.

Author information:

(1)McGill Centre for Bioinformatics, McGill University, Montreal, QC, Canada.

RNA molecules can achieve a broad range of regulatory functions through specific

structures that are in turn determined by their sequence. The prediction of

mutations changing the structural properties of RNA sequences (a.k.a. deleterious

mutations) is therefore useful for conducting mutagenesis experiments and

synthetic biology applications. While brute force approaches can be used to

analyze single-point mutations, this strategy does not scale well to multiple

mutations. In this article, we present corRna a web server for predicting the

multiple-point deleterious mutations in structural RNAs. corRna uses our

RNAmutants framework to efficiently explore the RNA mutational landscape. It also

enables users to apply search heuristics to improve the quality of the

predictions. We show that corRna predictions correlate with mutagenesis

experiments on the hepatitis C virus cis-acting replication element as well as

match the accuracy of previous approaches on a large test-set in a much lower

execution time. We illustrate these new perspectives offered by corRna by

predicting five-point deleterious mutations--an insight that could not be

achieved by previous methods. corRna is available at: http://corrna.cs.mcgill.ca.

DOI: 10.1093/nar/gkr358

PMCID: PMC3125766

PMID: 21596778 [Indexed for MEDLINE]

1669. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W215-22. doi:

10.1093/nar/gkr363. Epub 2011 May 18.

SDM--a server for predicting effects of mutations on protein stability and

malfunction.

Worth CL(1), Preissner R, Blundell TL.

Author information:

(1)Biochemistry Department, University of Cambridge, Cambridge CB2 1GA, UK.

The sheer volume of non-synonymous single nucleotide polymorphisms that have been

generated in recent years from projects such as the Human Genome Project, the

HapMap Project and Genome-Wide Association Studies means that it is not possible

to characterize all mutations experimentally on the gene products, i.e. elucidate

the effects of mutations on protein structure and function. However, automatic

methods that can predict the effects of mutations will allow a reduced set of

mutations to be studied. Site Directed Mutator (SDM) is a statistical potential

energy function that uses environment-specific amino-acid substitution

frequencies within homologous protein families to calculate a stability score,

which is analogous to the free energy difference between the wild-type and mutant

protein. Here, we present a web server for SDM

(http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php), which has obtained more than

10,000 submissions since being online in April 2008. To run SDM, users must

upload a wild-type structure and the position and amino acid type of the

mutation. The results returned include information about the local structural

environment of the wild-type and mutant residues, a stability score prediction

and prediction of disease association. Additionally, the wild-type and mutant

structures are displayed in a Jmol applet with the relevant residues highlighted.

DOI: 10.1093/nar/gkr363

PMCID: PMC3125769

PMID: 21593128 [Indexed for MEDLINE]

1670. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W29-37. doi: 10.1093/nar/gkr367.

Epub 2011 May 18.

HMMER web server: interactive sequence similarity searching.

Finn RD(1), Clements J, Eddy SR.

Author information:

(1)HHMI Janelia Farm Research Campus, 19700 Helix Drive, Ashburn, VA 20147, USA.

finnr@janelia.hhmi.org

HMMER is a software suite for protein sequence similarity searches using

probabilistic methods. Previously, HMMER has mainly been available only as a

computationally intensive UNIX command-line tool, restricting its use. Recent

advances in the software, HMMER3, have resulted in a 100-fold speed gain relative

to previous versions. It is now feasible to make efficient profile hidden Markov

model (profile HMM) searches via the web. A HMMER web server

(http://hmmer.janelia.org) has been designed and implemented such that most

protein database searches return within a few seconds. Methods are available for

searching either a single protein sequence, multiple protein sequence alignment

or profile HMM against a target sequence database, and for searching a protein

sequence against Pfam. The web server is designed to cater to a range of

different user expertise and accepts batch uploading of multiple queries at once.

All search methods are also available as RESTful web services, thereby allowing

them to be readily integrated as remotely executed tasks in locally scripted

workflows. We have focused on minimizing search times and the ability to rapidly

display tabular results, regardless of the number of matches found, developing

graphical summaries of the search results to provide quick, intuitive

appraisement of them.

DOI: 10.1093/nar/gkr367

PMCID: PMC3125773

PMID: 21593126 [Indexed for MEDLINE]

1671. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W79-85. doi: 10.1093/nar/gkr291.

Epub 2011 May 18.

Genome Surveyor 2.0: cis-regulatory analysis in Drosophila.

Kazemian M(1), Brodsky MH, Sinha S.

Author information:

(1)Department of Computer Science, University of Illinois at Urbana-Champaign,

Urbana, IL 61801, USA.

Genome Surveyor 2.0 is a web-based tool for discovery and analysis of

cis-regulatory elements in Drosophila, built on top of the GBrowse genome browser

for convenient visualization. Genome Surveyor was developed as a tool for

predicting transcription factor (TF) binding targets and cis-regulatory modules

(CRMs/enhancers), based on motifs representing experimentally determined DNA

binding specificities. Since its first publication, we have added substantial new

functionality (e.g. phylogenetic averaging of motif scores from multiple species,

and a novel CRM discovery technique), increased the number of supported motifs

about 4-fold (from ∼100 to ∼400), added provisions for evolutionary comparison

across many more Drosophila species (from 2 to 12), and improved the

user-interface. The server is free and open to all users, and there is no login

requirement. Address: http://veda.cs.uiuc.edu/gs.

DOI: 10.1093/nar/gkr291

PMCID: PMC3125742

PMID: 21593125 [Indexed for MEDLINE]

1672. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W430-6. doi: 10.1093/nar/gkr332.

Epub 2011 May 17.

ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression

data to discover direct/indirect targets of a transcription factor.

Qin J(1), Li MJ, Wang P, Zhang MQ, Wang J.

Author information:

(1)Department of Biochemistry, LKS Faculty of Medicine, The University of Hong

Kong, 21 Sassoon Road, Hong Kong SAR, China.

Chromatin immunoprecipitation (ChIP) coupled with high-throughput techniques

(ChIP-X), such as next generation sequencing (ChIP-Seq) and microarray

(ChIP-chip), has been successfully used to map active transcription factor

binding sites (TFBS) of a transcription factor (TF). The targeted genes can be

activated or suppressed by the TF, or are unresponsive to the TF. Microarray

technology has been used to measure the actual expression changes of thousands of

genes under the perturbation of a TF, but is unable to determine if the affected

genes are direct or indirect targets of the TF. Furthermore, both ChIP-X and

microarray methods produce a large number of false positives. Combining

microarray expression profiling and ChIP-X data allows more effective TFBS

analysis for studying the function of a TF. However, current web servers only

provide tools to analyze either ChIP-X or expression data, but not both. Here, we

present ChIP-Array, a web server that integrates ChIP-X and expression data from

human, mouse, yeast, fruit fly and Arabidopsis. This server will assist

biologists to detect direct and indirect target genes regulated by a TF of

interest and to aid in the functional characterization of the TF. ChIP-Array is

available at http://jjwanglab.hku.hk/ChIP-Array, with free access to academic

users.

DOI: 10.1093/nar/gkr332

PMCID: PMC3125757

PMID: 21586587 [Indexed for MEDLINE]

1673. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W18-23. doi: 10.1093/nar/gkr333.

Epub 2011 May 17.

iPBA: a tool for protein structure comparison using sequence alignment

strategies.

Gelly JC(1), Joseph AP, Srinivasan N, de Brevern AG.

Author information:

(1)INSERM, UMR-S 665, Dynamique des Structures et Interactions des Macromolécules

Biologiques, Université Paris Diderot-Paris 7, Institut National de la

Transfusion Sanguine, 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.

With the immense growth in the number of available protein structures, fast and

accurate structure comparison has been essential. We propose an efficient method

for structure comparison, based on a structural alphabet. Protein Blocks (PBs) is

a widely used structural alphabet with 16 pentapeptide conformations that can

fairly approximate a complete protein chain. Thus a 3D structure can be

translated into a 1D sequence of PBs. With a simple Needleman-Wunsch approach and

a raw PB substitution matrix, PB-based structural alignments were better than

many popular methods. iPBA web server presents an improved alignment approach

using (i) specialized PB Substitution Matrices (SM) and (ii) anchor-based

alignment methodology. With these developments, the quality of ∼88% of alignments

was improved. iPBA alignments were also better than DALI, MUSTANG and GANGSTA(+)

in >80% of the cases. The webserver is designed to for both pairwise comparisons

and database searches. Outputs are given as sequence alignment and superposed 3D

structures displayed using PyMol and Jmol. A local alignment option for detecting

subs-structural similarity is also embedded. As a fast and efficient

'sequence-based' structure comparison tool, we believe that it will be quite

useful to the scientific community. iPBA can be accessed at

http://www.dsimb.inserm.fr/dsimb\_tools/ipba/.

DOI: 10.1093/nar/gkr333

PMCID: PMC3125758

PMID: 21586582 [Indexed for MEDLINE]

1674. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W424-9. doi: 10.1093/nar/gkr359.

Epub 2011 May 16.

ResponseNet: revealing signaling and regulatory networks linking genetic and

transcriptomic screening data.

Lan A(1), Smoly IY, Rapaport G, Lindquist S, Fraenkel E, Yeger-Lotem E.

Author information:

(1)Department of Computer Science, Ben-Gurion University of The Negev, Beer-Sheva

84105, Israel.

Cellular response to stimuli is typically complex and involves both regulatory

and metabolic processes. Large-scale experimental efforts to identify components

of these processes often comprise of genetic screening and transcriptomic

profiling assays. We previously established that in yeast genetic screens tend to

identify response regulators, while transcriptomic profiling assays tend to

identify components of metabolic processes. ResponseNet is a network-optimization

approach that integrates the results from these assays with data of known

molecular interactions. Specifically, ResponseNet identifies a high-probability

sub-network, composed of signaling and regulatory molecular interaction paths,

through which putative response regulators may lead to the measured

transcriptomic changes. Computationally, this is achieved by formulating a

minimum-cost flow optimization problem and solving it efficiently using linear

programming tools. The ResponseNet web server offers a simple interface for

applying ResponseNet. Users can upload weighted lists of proteins and genes and

obtain a sparse, weighted, molecular interaction sub-network connecting their

data. The predicted sub-network and its gene ontology enrichment analysis are

presented graphically or as text. Consequently, the ResponseNet web server

enables researchers that were previously limited to separate analysis of their

distinct, large-scale experiments, to meaningfully integrate their data and

substantially expand their understanding of the underlying cellular response.

ResponseNet is available at http://bioinfo.bgu.ac.il/respnet.

DOI: 10.1093/nar/gkr359

PMCID: PMC3125767

PMID: 21576238 [Indexed for MEDLINE]

1675. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W288-92. doi:

10.1093/nar/gkr365. Epub 2011 May 16.

MetalDetector v2.0: predicting the geometry of metal binding sites from protein

sequence.

Passerini A(1), Lippi M, Frasconi P.

Author information:

(1)Dipartimento di Ingegneria e Scienza dell'Informazione, Università degli Studi

di Trento, Via Sommarive 14, 38123 Povo di Trento, Italy. passerini@disi.unitn.it

MetalDetector identifies CYS and HIS involved in transition metal protein binding

sites, starting from sequence alone. A major new feature of release 2.0 is the

ability to predict which residues are jointly involved in the coordination of the

same metal ion. The server is available at

http://metaldetector.dsi.unifi.it/v2.0/.

DOI: 10.1093/nar/gkr365

PMCID: PMC3125771

PMID: 21576237 [Indexed for MEDLINE]

1676. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W74-8. doi: 10.1093/nar/gkr355.

Epub 2011 May 16.

ConTra v2: a tool to identify transcription factor binding sites across species,

update 2011.

Broos S(1), Hulpiau P, Galle J, Hooghe B, Van Roy F, De Bleser P.

Author information:

(1)Department for Molecular Biomedical Research, VIB, Ghent University, Ghent,

Belgium. stefanb@dmbr.vib-ugent.be

Transcription factors are important gene regulators with distinctive roles in

development, cell signaling and cell cycling, and they have been associated with

many diseases. The ConTra v2 web server allows easy visualization and exploration

of predicted transcription factor binding sites in any genomic region surrounding

coding or non-coding genes. In this new version, users can choose from nine

reference organisms ranging from human to yeast. ConTra v2 can analyze promoter

regions, 5'-UTRs, 3'-UTRs and introns or any other genomic region of interest.

Hundreds of position weight matrices are available to choose from, but the user

can also upload any other matrices for detecting specific binding sites. A

typical analysis is run in four simple steps of choosing the gene, the

transcript, the region of interest and then selecting one or more transcription

factor binding sites. The ConTra v2 web server is freely available at

http://bioit.dmbr.ugent.be/contrav2/index.php.

DOI: 10.1093/nar/gkr355

PMCID: PMC3125763

PMID: 21576231 [Indexed for MEDLINE]

1677. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W229-34. doi:

10.1093/nar/gkr317. Epub 2011 May 16.

PRUNE and PROBE--two modular web services for protein-protein docking.

Mitra P(1), Pal D.

Author information:

(1)Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India.

The protein-protein docking programs typically perform four major tasks: (i)

generation of docking poses, (ii) selecting a subset of poses, (iii) their

structural refinement and (iv) scoring, ranking for the final assessment of the

true quaternary structure. Although the tasks can be integrated or performed in a

serial order, they are by nature modular, allowing an opportunity to substitute

one algorithm with another. We have implemented two modular web services, (i)

PRUNE: to select a subset of docking poses generated during sampling search

(http://pallab.serc.iisc.ernet.in/prune) and (ii) PROBE: to refine, score and

rank them (http://pallab.serc.iisc.ernet.in/probe). The former uses a new

interface area based edge-scoring function to eliminate >95% of the poses

generated during docking search. In contrast to other multi-parameter-based

screening functions, this single parameter based elimination reduces the

computational time significantly, in addition to increasing the chances of

selecting native-like models in the top rank list. The PROBE server performs

ranking of pruned poses, after structure refinement and scoring using a

regression model for geometric compatibility, and normalized interaction energy.

While web-service similar to PROBE is infrequent, no web-service akin to PRUNE

has been described before. Both the servers are publicly accessible and free for

use.

DOI: 10.1093/nar/gkr317

PMCID: PMC3125751

PMID: 21576226 [Indexed for MEDLINE]

1678. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W210-4. doi: 10.1093/nar/gkr352.

Epub 2011 May 16.

The FALC-Loop web server for protein loop modeling.

Ko J(1), Lee D, Park H, Coutsias EA, Lee J, Seok C.

Author information:

(1)Department of Chemistry, Seoul National University, Seoul 151-747, Korea.

The FALC-Loop web server provides an online interface for protein loop modeling

by employing an ab initio loop modeling method called FALC (fragment assembly and

analytical loop closure). The server may be used to construct loop regions in

homology modeling, to refine unreliable loop regions in experimental structures

or to model segments of designed sequences. The FALC method is computationally

less expensive than typical ab initio methods because the conformational search

space is effectively reduced by the use of fragments derived from a structure

database. The analytical loop closure algorithm allows efficient search for loop

conformations that fit into the protein framework starting from the

fragment-assembled structures. The FALC method shows prediction accuracy

comparable to other state-of-the-art loop modeling methods. Top-ranked model

structures can be visualized on the web server, and an ensemble of loop

structures can be downloaded for further analysis. The web server can be freely

accessed at http://falc-loop.seoklab.org/.

DOI: 10.1093/nar/gkr352

PMCID: PMC3125760

PMID: 21576220 [Indexed for MEDLINE]

1679. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W444-9. doi: 10.1093/nar/gkr321.

Epub 2011 May 16.

SNPsyn: detection and exploration of SNP-SNP interactions.

Curk T(1), Rot G, Zupan B.

Author information:

(1)Faculty of Computer and Information Science, University of Ljubljana, Trzaska

cesta 25, SI-1000 Ljubljana, Slovenia. tomaz.curk@fri.uni-lj.si

SNPsyn (http://snpsyn.biolab.si) is an interactive software tool for the

discovery of synergistic pairs of single nucleotide polymorphisms (SNPs) from

large genome-wide case-control association studies (GWAS) data on complex

diseases. Synergy among SNPs is estimated using an information-theoretic approach

called interaction analysis. SNPsyn is both a stand-alone C++/Flash application

and a web server. The computationally intensive part is implemented in C++ and

can run in parallel on a dedicated cluster or grid. The graphical user interface

is written in Adobe Flash Builder 4 and can run in most web browsers or as a

stand-alone application. The SNPsyn web server hosts the Flash application,

receives GWAS data submissions, invokes the interaction analysis and serves

result files. The user can explore details on identified synergistic pairs of

SNPs, perform gene set enrichment analysis and interact with the constructed SNP

synergy network.

DOI: 10.1093/nar/gkr321

PMCID: PMC3125755

PMID: 21576219 [Indexed for MEDLINE]

1680. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W381-4. doi: 10.1093/nar/gkr318.

Epub 2011 May 13.

PepServe: a web server for peptide analysis, clustering and visualization.

Alexandridou A(1), Dovrolis N, Tsangaris GT, Nikita K, Spyrou G.

Author information:

(1)Biomedical Research Foundation, Academy of Athens, 4 Soranou Ephessiou, 115 27

Athens, Greece.

Peptides, either as protein fragments or as naturally occurring entities are

characterized by their sequence and function features. Many times the researchers

need to massively manage peptide lists concerning protein identification,

biomarker discovery, bioactivity, immune response or other functionalities. We

present a web server that manages peptide lists in terms of feature analysis as

well as interactive clustering and visualization of the given peptides. PepServe

is a useful tool in the understanding of the peptide feature distribution among a

group of peptides. The PepServe web application is freely available at

http://bioserver-1.bioacademy.gr/Bioserver/PepServe/.

DOI: 10.1093/nar/gkr318

PMCID: PMC3125752

PMID: 21572105 [Indexed for MEDLINE]

1681. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W68-73. doi: 10.1093/nar/gkr316.

Epub 2011 May 10.

CURVES+ web server for analyzing and visualizing the helical, backbone and groove

parameters of nucleic acid structures.

Blanchet C(1), Pasi M, Zakrzewska K, Lavery R.

Author information:

(1)Université Lyon 1, CNRS, UMR 5086, Bases Moléculaires et Structurales des

Systèmes Infectieux, IBCP FR3302, 7 passage du Vercors, F-69367, France.

Curves+, a revised version of the Curves software for analyzing the conformation

of nucleic acid structures, is now available as a web server. This version, which

can be freely accessed at http://gbio-pbil.ibcp.fr/cgi/Curves\_plus/, allows the

user to upload a nucleic acid structure file, choose the nucleotides to be

analyzed and after optionally setting a number of input variables, view the

numerical and graphic results online or download files containing a set of

helical, backbone and groove parameters that fully describe the structure. PDB

format files are also provided for offline visualization of the helical axis and

groove geometry.

DOI: 10.1093/nar/gkr316

PMCID: PMC3125750

PMID: 21558323 [Indexed for MEDLINE]

1682. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W492-8. doi: 10.1093/nar/gkr299.

Epub 2011 May 10.

DRAR-CPI: a server for identifying drug repositioning potential and adverse drug

reactions via the chemical-protein interactome.

Luo H(1), Chen J, Shi L, Mikailov M, Zhu H, Wang K, He L, Yang L.

Author information:

(1)Bio-X Institutes, Key Laboratory for the Genetics of Developmental and

Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong

University, Shanghai, China.

Identifying new indications for existing drugs (drug repositioning) is an

efficient way of maximizing their potential. Adverse drug reaction (ADR) is one

of the leading causes of death among hospitalized patients. As both new

indications and ADRs are caused by unexpected chemical-protein interactions on

off-targets, it is reasonable to predict these interactions by mining the

chemical-protein interactome (CPI). Making such predictions has recently been

facilitated by a web server named DRAR-CPI. This server has a representative

collection of drug molecules and targetable human proteins built up from our work

in drug repositioning and ADR. When a user submits a molecule, the server will

give the positive or negative association scores between the user's molecule and

our library drugs based on their interaction profiles towards the targets. Users

can thus predict the indications or ADRs of their molecule based on the

association scores towards our library drugs. We have matched our predictions of

drug-drug associations with those predicted via gene-expression profiles,

achieving a matching rate as high as 74%. We have also successfully predicted the

connections between anti-psychotics and anti-infectives, indicating the

underlying relevance of anti-psychotics in the potential treatment of infections,

vice versa. This server is freely available at http://cpi.bio-x.cn/drar/.

DOI: 10.1093/nar/gkr299

PMCID: PMC3125745

PMID: 21558322 [Indexed for MEDLINE]

1683. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W13-7. doi: 10.1093/nar/gkr245.

Epub 2011 May 9.

T-Coffee: a web server for the multiple sequence alignment of protein and RNA

sequences using structural information and homology extension.

Di Tommaso P(1), Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang JM, Taly

JF, Notredame C.

Author information:

(1)Centre For Genomic Regulation (Pompeu Fabra University), Carrer del Doctor

Aiguader 88, 08003 Barcelona, Spain.

This article introduces a new interface for T-Coffee, a consistency-based

multiple sequence alignment program. This interface provides an easy and

intuitive access to the most popular functionality of the package. These include

the default T-Coffee mode for protein and nucleic acid sequences, the M-Coffee

mode that allows combining the output of any other aligners, and template-based

modes of T-Coffee that deliver high accuracy alignments while using structural or

homology derived templates. These three available template modes are Expresso for

the alignment of protein with a known 3D-Structure, R-Coffee to align RNA

sequences with conserved secondary structures and PSI-Coffee to accurately align

distantly related sequences using homology extension. The new server benefits

from recent improvements of the T-Coffee algorithm and can align up to 150

sequences as long as 10,000 residues and is available from both

http://www.tcoffee.org and its main mirror http://tcoffee.crg.cat.

DOI: 10.1093/nar/gkr245

PMCID: PMC3125728

PMID: 21558174 [Indexed for MEDLINE]

1684. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W362-7. doi: 10.1093/nar/gkr323.

Epub 2011 May 9.

NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity.

Röttig M(1), Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O.

Author information:

(1)Applied Bioinformatics, Center for Bioinformatics, Department of Computer

Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany.

roettig@informatik.uni-tuebingen.de

The products of many bacterial non-ribosomal peptide synthetases (NRPS) are

highly important secondary metabolites, including vancomycin and other

antibiotics. The ability to predict substrate specificity of newly detected NRPS

Adenylation (A-) domains by genome sequencing efforts is of great importance to

identify and annotate new gene clusters that produce secondary metabolites.

Prediction of A-domain specificity based on the sequence alone can be achieved

through sequence signatures or, more accurately, through machine learning

methods. We present an improved predictor, based on previous work

(NRPSpredictor), that predicts A-domain specificity using Support Vector Machines

on four hierarchical levels, ranging from gross physicochemical properties of an

A-domain's substrates down to single amino acid substrates. The three more

general levels are predicted with an F-measure better than 0.89 and the most

detailed level with an average F-measure of 0.80. We also modeled the

applicability domain of our predictor to estimate for new A-domains whether they

lie in the applicability domain. Finally, since there are also NRPS that play an

important role in natural products chemistry of fungi, such as peptaibols and

cephalosporins, we added a predictor for fungal A-domains, which predicts gross

physicochemical properties with an F-measure of 0.84. The service is available at

http://nrps.informatik.uni-tuebingen.de/.

DOI: 10.1093/nar/gkr323

PMCID: PMC3125756

PMID: 21558170 [Indexed for MEDLINE]

1685. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W145-8. doi: 10.1093/nar/gkr294.

Epub 2011 May 6.

DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and

bibliographic miRNA to disease association.

Maragkakis M(1), Vergoulis T, Alexiou P, Reczko M, Plomaritou K, Gousis M,

Kourtis K, Koziris N, Dalamagas T, Hatzigeorgiou AG.

Author information:

(1)Institute of Molecular Oncology, Biomedical Sciences Research Center Alexander

Fleming, 16672, Vari, Greece.

microRNAs (miRNAs) are small endogenous RNA molecules that are implicated in many

biological processes through post-transcriptional regulation of gene expression.

The DIANA-microT Web server provides a user-friendly interface for comprehensive

computational analysis of miRNA targets in human and mouse. The server has now

been extended to support predictions for two widely studied species: Drosophila

melanogaster and Caenorhabditis elegans. In the updated version, the Web server

enables the association of miRNAs to diseases through bibliographic analysis and

provides insights for the potential involvement of miRNAs in biological

processes. The nomenclature used to describe mature miRNAs along different

miRBase versions has been extensively analyzed, and the naming history of each

miRNA has been extracted. This enables the identification of miRNA publications

regardless of possible nomenclature changes. User interaction has been further

refined allowing users to save results that they wish to analyze further. A

connection to the UCSC genome browser is now provided, enabling users to easily

preview predicted binding sites in comparison to a wide array of genomic tracks,

such as single nucleotide polymorphisms. The Web server is publicly accessible in

www.microrna.gr/microT-v4.

DOI: 10.1093/nar/gkr294

PMCID: PMC3125744

PMID: 21551220 [Indexed for MEDLINE]

1686. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W375-80. doi:

10.1093/nar/gkr282. Epub 2011 May 4.

MemPype: a pipeline for the annotation of eukaryotic membrane proteins.

Pierleoni A(1), Indio V, Savojardo C, Fariselli P, Martelli PL, Casadio R.

Author information:

(1)Externautics s.p.a.-Bioinformatics, Via Fiorentina 1, 53100 Siena, Italy.

andrea.pierleoni@externautics.com

MemPype is a Python-based pipeline including previously published methods for the

prediction of signal peptides (SPEP), glycophosphatidylinositol (GPI) anchors

(PredGPI), all-alpha membrane topology (ENSEMBLE), and a recent method (MemLoci)

that specifically discriminates the localization of eukaryotic membrane proteins

in: 'cell membrane', 'internal membranes', 'organelle membranes'. MemLoci scores

with accuracy of 70% and generalized correlation coefficient (GCC) of 0.50 on a

rigorous homology-unbiased validation set and overpasses other predictors for

subcellular localization. The annotation process is based both on inheritance

through homology and computational methods. Each submitted protein first

retrieves, when available, up to 25 similar proteins (with sequence identity ≥50%

and alignment coverage ≥50% on both sequences). This helps the identification of

membrane-associated proteins and detailed localization tags. Each protein is also

filtered for the presence of a GPI anchor [0.8% false positive rate (FPR)]. A

positive score of GPI anchor prediction labels the sequence as exposed to 'Cell

surface'. Concomitantly the sequence is analysed for the presence of a signal

peptide and classified with MemLoci into one of three discriminated classes.

Finally the sequence is filtered for predicting its putative all-alpha protein

membrane topology (FPR <1%). The web server is available at:

http://mu2py.biocomp.unibo.it/mempype.

DOI: 10.1093/nar/gkr282

PMCID: PMC3125734

PMID: 21543452 [Indexed for MEDLINE]

1687. Nucleic Acids Res. 2011 Jul;39(13):e87. doi: 10.1093/nar/gkr251. Epub 2011 May 4.

NOA: a novel Network Ontology Analysis method.

Wang J(1), Huang Q, Liu ZP, Wang Y, Wu LY, Chen L, Zhang XS.

Author information:

(1)Key Laboratory of Management, Decision and Information Systems, Academy of

Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190,

Japan.

Gene ontology analysis has become a popular and important tool in bioinformatics

study, and current ontology analyses are mainly conducted in individual gene or a

gene list. However, recent molecular network analysis reveals that the same list

of genes with different interactions may perform different functions. Therefore,

it is necessary to consider molecular interactions to correctly and specifically

annotate biological networks. Here, we propose a novel Network Ontology Analysis

(NOA) method to perform gene ontology enrichment analysis on biological networks.

Specifically, NOA first defines link ontology that assigns functions to

interactions based on the known annotations of joint genes via optimizing two

novel indexes 'Coverage' and 'Diversity'. Then, NOA generates two alternative

reference sets to statistically rank the enriched functional terms for a given

biological network. We compare NOA with traditional enrichment analysis methods

in several biological networks, and find that: (i) NOA can capture the change of

functions not only in dynamic transcription regulatory networks but also in

rewiring protein interaction networks while the traditional methods cannot and

(ii) NOA can find more relevant and specific functions than traditional methods

in different types of static networks. Furthermore, a freely accessible web

server for NOA has been developed at http://www.aporc.org/noa/.

DOI: 10.1093/nar/gkr251

PMCID: PMC3141273

PMID: 21543451 [Indexed for MEDLINE]

1688. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W479-85. doi:

10.1093/nar/gkr243. Epub 2011 Apr 29.

PhyleasProg: a user-oriented web server for wide evolutionary analyses.

Busset J(1), Cabau C, Meslin C, Pascal G.

Author information:

(1)INRA, UMR85, Physiologie de la Reproduction et des Comportements, F-37380

Nouzilly, France.

Evolutionary analyses of biological data are becoming a prerequisite in many

fields of biology. At a time of high-throughput data analysis, phylogenetics is

often a necessary complementary tool for biologists to understand, compare and

identify the functions of sequences. But available bioinformatics tools are

frequently not easy for non-specialists to use. We developed PhyleasProg

(http://phyleasprog.inra.fr), a user-friendly web server as a turnkey tool

dedicated to evolutionary analyses. PhyleasProg can help biologists with little

experience in evolutionary methodologies by analysing their data in a simple and

robust way, using methods corresponding to robust standards. Via a very intuitive

web interface, users only need to enter a list of Ensembl protein IDs and a list

of species as inputs. After dynamic computations, users have access to

phylogenetic trees, positive/purifying selection data (on site and branch-site

models), with a display of these results on the protein sequence and on a 3D

structure model, and the synteny environment of related genes. This connection

between different domains of phylogenetics opens the way to new biological

analyses for the discovery of the function and structure of proteins.

DOI: 10.1093/nar/gkr243

PMCID: PMC3125726

PMID: 21531699 [Indexed for MEDLINE]

1689. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W50-5. doi: 10.1093/nar/gkr249.

Epub 2011 Apr 22.

WebFR3D--a server for finding, aligning and analyzing recurrent RNA 3D motifs.

Petrov AI(1), Zirbel CL, Leontis NB.

Author information:

(1)Department of Biological Sciences, Bowling Green State University, Bowling

Green, OH, USA.

WebFR3D is the on-line version of 'Find RNA 3D' (FR3D), a program for annotating

atomic-resolution RNA 3D structure files and searching them efficiently to locate

and compare RNA 3D structural motifs. WebFR3D provides on-line access to the

central features of FR3D, including geometric and symbolic search modes, without

need for installing programs or downloading and maintaining 3D structure data

locally. In geometric search mode, WebFR3D finds all motifs similar to a

user-specified query structure. In symbolic search mode, WebFR3D finds all sets

of nucleotides making user-specified interactions. In both modes, users can

specify sequence, sequence-continuity, base pairing, base-stacking and other

constraints on nucleotides and their interactions. WebFR3D can be used to locate

hairpin, internal or junction loops, list all base pairs or other interactions,

or find instances of recurrent RNA 3D motifs (such as sarcin-ricin and kink-turn

internal loops or T- and GNRA hairpin loops) in any PDB file or across a whole

set of 3D structure files. The output page provides facilities for comparing the

instances returned by the search by superposition of the 3D structures and the

alignment of their sequences annotated with pairwise interactions. WebFR3D is

available at http://rna.bgsu.edu/webfr3d.

DOI: 10.1093/nar/gkr249

PMCID: PMC3125732

PMID: 21515634 [Indexed for MEDLINE]

1690. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W132-8. doi: 10.1093/nar/gkr247.

Epub 2011 Apr 22.

miRanalyzer: an update on the detection and analysis of microRNAs in

high-throughput sequencing experiments.

Hackenberg M(1), Rodríguez-Ezpeleta N, Aransay AM.

Author information:

(1)Departmento de Genética, Facultad de Ciencias, Universidad de Granada, Campus

de Fuentenueva s/n, 18071-Granada, Spain. mlhack@gmail.com

We present a new version of miRanalyzer, a web server and stand-alone tool for

the detection of known and prediction of new microRNAs in high-throughput

sequencing experiments. The new version has been notably improved regarding

speed, scope and available features. Alignments are now based on the ultrafast

short-read aligner Bowtie (granting also colour space support, allowing

mismatches and improving speed) and 31 genomes, including 6 plant genomes, can

now be analysed (previous version contained only 7). Differences between plant

and animal microRNAs have been taken into account for the prediction models and

differential expression of both, known and predicted microRNAs, between two

conditions can be calculated. Additionally, consensus sequences of predicted

mature and precursor microRNAs can be obtained from multiple samples, which

increases the reliability of the predicted microRNAs. Finally, a stand-alone

version of the miRanalyzer that is based on a local and easily customized

database is also available; this allows the user to have more control on certain

parameters as well as to use specific data such as unpublished assemblies or

other libraries that are not available in the web server. miRanalyzer is

available at http://bioinfo2.ugr.es/miRanalyzer/miRanalyzer.php.

DOI: 10.1093/nar/gkr247

PMCID: PMC3125730

PMID: 21515631 [Indexed for MEDLINE]

1691. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W92-9. doi: 10.1093/nar/gkr207.

Epub 2011 Apr 7.

The RNAmute web server for the mutational analysis of RNA secondary structures.

Churkin A(1), Gabdank I, Barash D.

Author information:

(1)Department of Computer Science, Ben-Gurion University, Beer-Sheva 84105,

Israel.

RNA mutational analysis at the secondary-structure level can be useful to a

wide-range of biological applications. It can be used to predict an optimal site

for performing a nucleotide mutation at the single molecular level, as well as to

analyze basic phenomena at the systems level. For the former, as more sequence

modification experiments are performed that include site-directed mutagenesis to

find and explore functional motifs in RNAs, a pre-processing step that helps

guide in planning the experiment becomes vital. For the latter, mutations are

generally accepted as a central mechanism by which evolution occurs, and

mutational analysis relating to structure should gain a better understanding of

system functionality and evolution. In the past several years, the program

RNAmute that is structure based and relies on RNA secondary-structure prediction

has been developed for assisting in RNA mutational analysis. It has been extended

from single-point mutations to treat multiple-point mutations efficiently by

initially calculating all suboptimal solutions, after which only the mutations

that stabilize the suboptimal solutions and destabilize the optimal one are

considered as candidates for being deleterious. The RNAmute web server for

mutational analysis is available at http://www.cs.bgu.ac.il/~xrnamute/XRNAmute.

DOI: 10.1093/nar/gkr207

PMCID: PMC3125725

PMID: 21478166 [Indexed for MEDLINE]

1692. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W171-6. doi: 10.1093/nar/gkr184.

Epub 2011 Mar 31.

The IntFOLD server: an integrated web resource for protein fold recognition, 3D

model quality assessment, intrinsic disorder prediction, domain prediction and

ligand binding site prediction.

Roche DB(1), Buenavista MT, Tetchner SJ, McGuffin LJ.

Author information:

(1)School of Biological Sciences, University of Reading, Whiteknights, Reading

RG6 6AS, UK.

The IntFOLD server is a novel independent server that integrates several cutting

edge methods for the prediction of structure and function from sequence. Our

guiding principles behind the server development were as follows: (i) to provide

a simple unified resource that makes our prediction software accessible to all

and (ii) to produce integrated output for predictions that can be easily

interpreted. The output for predictions is presented as a simple table that

summarizes all results graphically via plots and annotated 3D models. The raw

machine readable data files for each set of predictions are also provided for

developers, which comply with the Critical Assessment of Methods for Protein

Structure Prediction (CASP) data standards. The server comprises an integrated

suite of five novel methods: nFOLD4, for tertiary structure prediction; ModFOLD

3.0, for model quality assessment; DISOclust 2.0, for disorder prediction;

DomFOLD 2.0 for domain prediction; and FunFOLD 1.0, for ligand binding site

prediction. Predictions from the IntFOLD server were found to be competitive in

several categories in the recent CASP9 experiment. The IntFOLD server is

available at the following web site: http://www.reading.ac.uk/bioinf/IntFOLD/.

DOI: 10.1093/nar/gkr184

PMCID: PMC3125722

PMID: 21459847 [Indexed for MEDLINE]

1693. OMICS. 2011 Jul-Aug;15(7-8):483-94. doi: 10.1089/omi.2010.0066. Epub 2011 Jun 23.

A topology-preserving selection and clustering approach to multidimensional

biological data.

Fang H(1), Du Y, Xia L, Li J, Zhang J, Wang K.

Author information:

(1)State Key Laboratory of Medical Genomics, Sino-French Research Center for Life

Sciences and Genomics, Ruijin Hospital affiliated to Shanghai Jiao Tong

University School of Medicine (SJTU-SM), People's Republic of China.

Multidimensional genome-wide data (e.g., gene expression microarray data) provide

rich information and widespread applications in integrative biology. However,

little attention has been paid to the inherent relationships within these natural

data. By simply viewing multidimensional microarray data scattered over

hyperspace, the spatial properties (topological structure) of the data clouds may

reveal the underlying relationships. Based on this idea, we herein make

analytical improvements by introducing a topology-preserving selection and

clustering (TPSC) approach to complex large-scale microarray data. Specifically,

the integration of self-organizing map (SOM) and singular value decomposition

allows genome-wide selection on sound foundations of statistical inference.

Moreover, this approach is complemented with an SOM-based two-phase gene

clustering procedure, allowing the topology-preserving identification of gene

clusters. These gene clusters with highly similar expression patterns can

facilitate many aspects of biological interpretations in terms of functional and

regulatory relevance. As demonstrated by processing large and complex datasets of

the human cell cycle, stress responses, and host cell responses to pathogen

infection, our proposed method can yield better characteristic features from the

whole datasets compared to conventional routines. We hence conclude that the

topology-preserving selection and clustering without a priori assumption on data

structure allow the in-depth mining of biological information in a more accurate

and unbiased manner. A Web server ( http://www.cs.bris.ac.uk/∼hfang/TPSC )

hosting a MATLAB package that implements the methodology is freely available to

both academic and nonacademic users. These advances will expand the scope of

omics applications.

DOI: 10.1089/omi.2010.0066

PMID: 21699401 [Indexed for MEDLINE]

1694. BMC Syst Biol. 2011 Jun 20;5 Suppl 1:S7. doi: 10.1186/1752-0509-5-S1-S7.

MetaDBSite: a meta approach to improve protein DNA-binding sites prediction.

Si J(1), Zhang Z, Lin B, Schroeder M, Huang B.

Author information:

(1)Systems Biology Division, Zhejiang-California International NanoSystems

Institute, Zhejiang University, Kaixuan Road 268, Hangzhou, China.

BACKGROUND: Protein-DNA interactions play an important role in many fundamental

biological activities such as DNA replication, transcription and repair.

Identification of amino acid residues involved in DNA binding site is critical

for understanding of the mechanism of gene regulations. In the last decade, there

have been a number of computational approaches developed to predict protein-DNA

binding sites based on protein sequence and/or structural information.

RESULTS: In this article, we present metaDBSite, a meta web server to predict

DNA-binding residues for DNA-binding proteins. MetaDBSite integrates the

prediction results from six available online web servers: DISIS, DNABindR, BindN,

BindN-rf, DP-Bind and DBS-PRED and it solely uses sequence information of

proteins. A large dataset of DNA-binding proteins is constructed from the Protein

Data Bank and it serves as a gold-standard benchmark to evaluate the metaDBSite

approach and the other six predictors.

CONCLUSIONS: The comparison results show that metaDBSite outperforms single

individual approach. We believe that metaDBSite will become a useful and

integrative tool for protein DNA-binding residues prediction. The MetaDBSite

web-server is freely available at

http://projects.biotec.tu-dresden.de/metadbsite/ and

http://sysbio.zju.edu.cn/metadbsite.

DOI: 10.1186/1752-0509-5-S1-S7

PMCID: PMC3121123

PMID: 21689482 [Indexed for MEDLINE]

1695. BMC Bioinformatics. 2011 Jun 17;12:245. doi: 10.1186/1471-2105-12-245.

In-silico prediction of disorder content using hybrid sequence representation.

Mizianty MJ(1), Zhang T, Xue B, Zhou Y, Dunker AK, Uversky VN, Kurgan L.

Author information:

(1)Department of Electrical and Computer Engineering, University of Alberta,

Edmonton, Alberta T6G 2V4, Canada.

BACKGROUND: Intrinsically disordered proteins play important roles in various

cellular activities and their prevalence was implicated in a number of human

diseases. The knowledge of the content of the intrinsic disorder in proteins is

useful for a variety of studies including estimation of the abundance of disorder

in protein families, classes, and complete proteomes, and for the analysis of

disorder-related protein functions. The above investigations currently utilize

the disorder content derived from the per-residue disorder predictions. We show

that these predictions may over-or under-predict the overall amount of disorder,

which motivates development of novel tools for direct and accurate sequence-based

prediction of the disorder content.

RESULTS: We hypothesize that sequence-level aggregation of input information may

provide more accurate content prediction when compared with the content extracted

from the local window-based residue-level disorder predictors. We propose a novel

predictor, DisCon, that takes advantage of a small set of 29 custom-designed

descriptors that aggregate and hybridize information concerning sequence,

evolutionary profiles, and predicted secondary structure, solvent accessibility,

flexibility, and annotation of globular domains. Using these descriptors and a

ridge regression model, DisCon predicts the content with low, 0.05, mean squared

error and high, 0.68, Pearson correlation. This is a statistically significant

improvement over the content computed from outputs of ten modern disorder

predictors on a test dataset with proteins that share low sequence identity with

the training sequences. The proposed predictive model is analyzed to discuss

factors related to the prediction of the disorder content.

CONCLUSIONS: DisCon is a high-quality alternative for high-throughput annotation

of the disorder content. We also empirically demonstrate that the DisCon's

predictions can be used to improve binary annotations of the disordered residues

from the real-value disorder propensities generated by current residue-level

disorder predictors. The web server that implements the DisCon is available at

http://biomine.ece.ualberta.ca/DisCon/.

DOI: 10.1186/1471-2105-12-245

PMCID: PMC3212983

PMID: 21682902 [Indexed for MEDLINE]

1696. BMC Bioinformatics. 2011 Jun 17;12:244. doi: 10.1186/1471-2105-12-244.

HomPPI: a class of sequence homology based protein-protein interface prediction

methods.

Xue LC(1), Dobbs D, Honavar V.

Author information:

(1)Department of Computer Science, Iowa State University, Ames, IA 50011, USA.

me.lixue@gmail.com

BACKGROUND: Although homology-based methods are among the most widely used

methods for predicting the structure and function of proteins, the question as to

whether interface sequence conservation can be effectively exploited in

predicting protein-protein interfaces has been a subject of debate.

RESULTS: We studied more than 300,000 pair-wise alignments of protein sequences

from structurally characterized protein complexes, including both obligate and

transient complexes. We identified sequence similarity criteria required for

accurate homology-based inference of interface residues in a query protein

sequence.Based on these analyses, we developed HomPPI, a class of sequence

homology-based methods for predicting protein-protein interface residues. We

present two variants of HomPPI: (i) NPS-HomPPI (Non partner-specific HomPPI),

which can be used to predict interface residues of a query protein in the absence

of knowledge of the interaction partner; and (ii) PS-HomPPI (Partner-specific

HomPPI), which can be used to predict the interface residues of a query protein

with a specific target protein.Our experiments on a benchmark dataset of obligate

homodimeric complexes show that NPS-HomPPI can reliably predict protein-protein

interface residues in a given protein, with an average correlation coefficient

(CC) of 0.76, sensitivity of 0.83, and specificity of 0.78, when sequence

homologs of the query protein can be reliably identified. NPS-HomPPI also

reliably predicts the interface residues of intrinsically disordered proteins.

Our experiments suggest that NPS-HomPPI is competitive with several

state-of-the-art interface prediction servers including those that exploit the

structure of the query proteins. The partner-specific classifier, PS-HomPPI can,

on a large dataset of transient complexes, predict the interface residues of a

query protein with a specific target, with a CC of 0.65, sensitivity of 0.69, and

specificity of 0.70, when homologs of both the query and the target can be

reliably identified. The HomPPI web server is available at

http://homppi.cs.iastate.edu/.

CONCLUSIONS: Sequence homology-based methods offer a class of computationally

efficient and reliable approaches for predicting the protein-protein interface

residues that participate in either obligate or transient interactions. For query

proteins involved in transient interactions, the reliability of interface residue

prediction can be improved by exploiting knowledge of putative interaction

partners.

DOI: 10.1186/1471-2105-12-244

PMCID: PMC3213298

PMID: 21682895 [Indexed for MEDLINE]

1697. Bioinformatics. 2011 Jun 15;27(12):1715-6. doi: 10.1093/bioinformatics/btr268.

Epub 2011 May 5.

APOLLO: a quality assessment service for single and multiple protein models.

Wang Z(1), Eickholt J, Cheng J.

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, MO 65211,

USA.

SUMMARY: We built a web server named APOLLO, which can evaluate the absolute

global and local qualities of a single protein model using machine learning

methods or the global and local qualities of a pool of models using a pair-wise

comparison approach. Based on our evaluations on 107 CASP9 (Critical Assessment

of Techniques for Protein Structure Prediction) targets, the predicted quality

scores generated from our machine learning and pair-wise methods have an average

per-target correlation of 0.671 and 0.917, respectively, with the true model

quality scores. Based on our test on 92 CASP9 targets, our predicted absolute

local qualities have an average difference of 2.60 Å with the actual distances to

native structure.

AVAILABILITY: http://sysbio.rnet.missouri.edu/apollo/. Single and pair-wise

global quality assessment software is also available at the site.

DOI: 10.1093/bioinformatics/btr268

PMCID: PMC3106203

PMID: 21546397 [Indexed for MEDLINE]

1698. Bioinformatics. 2011 Jun 15;27(12):1630-6. doi: 10.1093/bioinformatics/btr234.

Epub 2011 Apr 19.

A reference dataset for the analyses of membrane protein secondary structures and

transmembrane residues using circular dichroism spectroscopy.

Abdul-Gader A(1), Miles AJ, Wallace BA.

Author information:

(1)Department of Crystallography, Institute of Structural and Molecular Biology,

Birkbeck College, University of London, London WC1E 7HX, UK.

MOTIVATION: Empirical analyses of protein secondary structures based on circular

dichroism (CD) and synchrotron radiation circular dichroism (SRCD) spectroscopic

data rely on the availability of reference datasets comprised of spectra of

relevant proteins, whose crystal structures have been determined. Datasets

comprised of only soluble proteins have not proven suitable for analysing the

spectra of membrane proteins.

RESULTS: A new reference dataset, MP180, has been created containing the spectra

of 30 membrane proteins encompassing the secondary structure and fold space

covered by all known membrane protein structures. In addition a mixed soluble and

membrane protein dataset, SMP180, has been created, which includes 98 soluble

protein spectra (SP) plus the MP180 spectra. Calculations of both membrane and

soluble protein secondary structures using SMP180 are significantly improved with

respect to those produced, using soluble protein-only datasets. The SMP180

dataset also enables determination of the percentage of transmembrane residues,

thus enhancing the information previously obtainable from CD spectroscopy.

AVAILABILITY AND IMPLEMENTATION: Reference dataset online at the DichroWeb

analysis server (http://dichroweb.cryst.bbk.ac.uk); individual protein spectra in

the Protein Circular Dichroism Data Bank (http://pcddb.cryst.bbk.ac.uk).

DOI: 10.1093/bioinformatics/btr234

PMID: 21505036 [Indexed for MEDLINE]

1699. Bioinformatics. 2011 Jun 15;27(12):1734-5. doi: 10.1093/bioinformatics/btr181.

Epub 2011 Apr 14.

'Sciencenet'--towards a global search and share engine for all scientific

knowledge.

Lütjohann DS(1), Shah AH, Christen MP, Richter F, Knese K, Liebel U.

Author information:

(1)Institute of Organic Chemistry, Karlsruhe, Germany.

SUMMARY: Modern biological experiments create vast amounts of data which are

geographically distributed. These datasets consist of petabytes of raw data and

billions of documents. Yet to the best of our knowledge, a search engine

technology that searches and cross-links all different data types in life

sciences does not exist. We have developed a prototype distributed scientific

search engine technology, 'Sciencenet', which facilitates rapid searching over

this large data space. By 'bringing the search engine to the data', we do not

require server farms. This platform also allows users to contribute to the search

index and publish their large-scale data to support e-Science. Furthermore, a

community-driven method guarantees that only scientific content is crawled and

presented. Our peer-to-peer approach is sufficiently scalable for the science web

without performance or capacity tradeoff.

AVAILABILITY AND IMPLEMENTATION: The free to use search portal web page and the

downloadable client are accessible at: http://sciencenet.kit.edu. The web portal

for index administration is implemented in ASP.NET, the 'AskMe' experiment

publisher is written in Python 2.7, and the backend 'YaCy' search engine is based

on Java 1.6.

DOI: 10.1093/bioinformatics/btr181

PMCID: PMC3106183

PMID: 21493657 [Indexed for MEDLINE]

1700. Bioinformatics. 2011 Jun 15;27(12):1725-6. doi: 10.1093/bioinformatics/btr195.

Epub 2011 Apr 12.

Cobweb: a Java applet for network exploration and visualisation.

von Eichborn J(1), Bourne PE, Preissner R.

Author information:

(1)Structural Bioinformatics Group, Institute for Physiology, Charité -

Universitätsmedizin Berlin, Lindenberger Weg 80, 13125 Berlin, Germany.

joachim.eichborn@charite.de

SUMMARY: Cobweb is a Java applet for real-time network visualization; its

strength lies in enabling the interactive exploration of networks. Therefore, it

allows new nodes to be interactively added to a network by querying a database on

a server. The network constantly rearranges to provide the most meaningful

topological view.

AVAILABILITY: Cobweb is available under the GPLv3 and may be freely downloaded at

http://bioinformatics.charite.de/cobweb.

DOI: 10.1093/bioinformatics/btr195

PMID: 21486937 [Indexed for MEDLINE]

1701. BMC Res Notes. 2011 Jun 11;4:181. doi: 10.1186/1756-0500-4-181.

FARO server: Meta-analysis of gene expression by matching gene expression

signatures to a compendium of public gene expression data.

Manijak MP(1), Nielsen HB.

Author information:

(1)Center for Biological Sequence Analysis, Department of Systems Biology,

Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby,

Denmark. hbjorn@cbs.dtu.dk.

BACKGROUND: Although, systematic analysis of gene annotation is a powerful tool

for interpreting gene expression data, it sometimes is blurred by incomplete gene

annotation, missing expression response of key genes and secondary gene

expression responses. These shortcomings may be partially circumvented by instead

matching gene expression signatures to signatures of other experiments.

FINDINGS: To facilitate this we present the Functional Association Response by

Overlap (FARO) server, that match input signatures to a compendium of 242 gene

expression signatures, extracted from more than 1700 Arabidopsis microarray

experiments.

CONCLUSIONS: Hereby we present a publicly available tool for robust

characterization of Arabidopsis gene expression experiments which can point to

similar experimental factors in other experiments. The server is available at

http://www.cbs.dtu.dk/services/faro/.

DOI: 10.1186/1756-0500-4-181

PMCID: PMC3127962

PMID: 21663689

1702. Bioinformatics. 2011 Jun 1;27(11):1546-54. doi: 10.1093/bioinformatics/btr161.

Epub 2011 Apr 5.

Comprehensive and relaxed search for oligonucleotide signatures in hierarchically

clustered sequence datasets.

Bader KC(1), Grothoff C, Meier H.

Author information:

(1)Services Department of Informatics, Technische Universität München,

Boltzmannstrasse 3, 85748 Garching, Germany.

MOTIVATION: PCR, hybridization, DNA sequencing and other important methods in

molecular diagnostics rely on both sequence-specific and sequence group-specific

oligonucleotide primers and probes. Their design depends on the identification of

oligonucleotide signatures in whole genome or marker gene sequences. Although

genome and gene databases are generally available and regularly updated,

collections of valuable signatures are rare. Even for single requests, the search

for signatures becomes computationally expensive when working with large

collections of target (and non-target) sequences. Moreover, with growing dataset

sizes, the chance of finding exact group-matching signatures decreases,

necessitating the application of relaxed search methods. The resultant

substantial increase in complexity is exacerbated by the dearth of algorithms

able to solve these problems efficiently.

RESULTS: We have developed CaSSiS, a fast and scalable method for computing

comprehensive collections of sequence- and sequence group-specific

oligonucleotide signatures from large sets of hierarchically clustered nucleic

acid sequence data. Based on the ARB Positional Tree (PT-)Server and a newly

developed BGRT data structure, CaSSiS not only determines sequence-specific

signatures and perfect group-covering signatures for every node within the

cluster (i.e. target groups), but also signatures with maximal group coverage

(sensitivity) within a user-defined range of non-target hits (specificity) for

groups lacking a perfect common signature. An upper limit of tolerated mismatches

within the target group, as well as the minimum number of mismatches with

non-target sequences, can be predefined. Test runs with one of the largest

phylogenetic gene sequence datasets available indicate good runtime and memory

performance, and in silico spot tests have shown the usefulness of the resulting

signature sequences as blueprints for group-specific oligonucleotide probes.

AVAILABILITY: Software and Supplementary Material are available at

http://cassis.in.tum.de/.

DOI: 10.1093/bioinformatics/btr161

PMID: 21471017 [Indexed for MEDLINE]

1703. BMC Genomics. 2011 Jun 1;12:279. doi: 10.1186/1471-2164-12-279.

Transcriptome map of plant mitochondria reveals islands of unexpected transcribed

regions.

Fujii S(1), Toda T, Kikuchi S, Suzuki R, Yokoyama K, Tsuchida H, Yano K, Toriyama

K.

Author information:

(1)Laboratory of Environmental Biotechnology, Graduate School of Agricultural

Science, Tohoku University, Sendai 981-8555, Japan.

BACKGROUND: Plant mitochondria contain a relatively large amount of genetic

information, suggesting that their functional regulation may not be as

straightforward as that of metazoans. We used a genomic tiling array to draw a

transcriptomic atlas of Oryza sativa japonica (rice) mitochondria, which was

predicted to be approximately 490-kb long.

RESULTS: Whereas statistical analysis verified the transcription of all

previously known functional genes such as the ones related to oxidative

phosphorylation, a similar extent of RNA expression was frequently observed in

the inter-genic regions where none of the previously annotated genes are located.

The newly identified open reading frames (ORFs) predicted in these transcribed

inter-genic regions were generally not conserved among flowering plant species,

suggesting that these ORFs did not play a role in mitochondrial principal

functions. We also identified two partial fragments of retrotransposon sequences

as being transcribed in rice mitochondria.

CONCLUSION: The present study indicated the previously unexpected complexity of

plant mitochondrial RNA metabolism. Our transcriptomic data (Oryza sativa

Mitochondrial rna Expression Server: OsMES) is publicly accessible at

[http://bioinf.mind.meiji.ac.jp/cgi-bin/gbrowse/OsMes/#search].

DOI: 10.1186/1471-2164-12-279

PMCID: PMC3121727

PMID: 21627843 [Indexed for MEDLINE]

1704. J Comput Aided Mol Des. 2011 Jun;25(6):525-31. doi: 10.1007/s10822-011-9438-9.

Epub 2011 Jun 7.

iScreen: world's first cloud-computing web server for virtual screening and de

novo drug design based on TCM database@Taiwan.

Tsai TY(1), Chang KW, Chen CY.

Author information:

(1)Laboratory of Computational and Systems Biology, School of Chinese Medicine,

China Medical University, Taichung, 40402, Taiwan.

The rapidly advancing researches on traditional Chinese medicine (TCM) have

greatly intrigued pharmaceutical industries worldwide. To take initiative in the

next generation of drug development, we constructed a cloud-computing system for

TCM intelligent screening system (iScreen) based on TCM Database@Taiwan. iScreen

is compacted web server for TCM docking and followed by customized de novo drug

design. We further implemented a protein preparation tool that both extract

protein of interest from a raw input file and estimate the size of ligand bind

site. In addition, iScreen is designed in user-friendly graphic interface for

users who have less experience with the command line systems. For customized

docking, multiple docking services, including standard, in-water, pH environment,

and flexible docking modes are implemented. Users can download first 200 TCM

compounds of best docking results. For TCM de novo drug design, iScreen provides

multiple molecular descriptors for a user's interest. iScreen is the world's

first web server that employs world's largest TCM database for virtual screening

and de novo drug design. We believe our web server can lead TCM research to a new

era of drug development. The TCM docking and screening server is available at

http://iScreen.cmu.edu.tw/.

DOI: 10.1007/s10822-011-9438-9

PMID: 21647737 [Indexed for MEDLINE]

1705. Protein Pept Lett. 2011 Jun;18(6):573-87.

An efficient support vector machine approach for identifying protein

S-nitrosylation sites.

Li YX(1), Shao YH, Jing L, Deng NY.

Author information:

(1)College of Science, China Agricultural University, Beijing 100083, China.

Protein S-nitrosylation plays a key and specific role in many cellular processes.

Detecting possible S-nitrosylated substrates and their corresponding exact sites

is crucial for studying the mechanisms of these biological processes. Comparing

with the expensive and time-consuming biochemical experiments, the computational

methods are attracting considerable attention due to their convenience and fast

speed. Although some computational models have been developed to predict

S-nitrosylation sites, their accuracy is still low. In this work,we incorporate

support vector machine to predict protein S-nitrosylation sites. After a careful

evaluation of six encoding schemes, we propose a new efficient predictor,

CPR-SNO, using the coupling patterns based encoding scheme. The performance of

our CPR-SNO is measured with the area under the ROC curve (AUC) of 0.8289 in

10-fold cross validation experiments, which is significantly better than the

existing best method GPS-SNO 1.0's 0.685 performance. In further annotating

large-scale potential S-nitrosylated substrates, CPR-SNO also presents an

encouraging predictive performance. These results indicate that CPR-SNO can be

used as a competitive protein S-nitrosylation sites predictor to the biological

community. Our CPR-SNO has been implemented as a web server and is available at

http://math.cau.edu.cn/CPR -SNO/CPR-SNO.html.

PMID: 21271979 [Indexed for MEDLINE]

1706. Proteins. 2011 Jun;79(6):1952-63. doi: 10.1002/prot.23020. Epub 2011 Apr 12.

Structure-based identification of catalytic residues.

Yahalom R(1), Reshef D, Wiener A, Frankel S, Kalisman N, Lerner B, Keasar C.

Author information:

(1)Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva

84105, Israel.

The identification of catalytic residues is an essential step in functional

characterization of enzymes. We present a purely structural approach to this

problem, which is motivated by the difficulty of evolution-based methods to

annotate structural genomics targets that have few or no homologs in the

databases. Our approach combines a state-of-the-art support vector machine (SVM)

classifier with novel structural features that augment structural clues by

spatial averaging and Z scoring. Special attention is paid to the class imbalance

problem that stems from the overwhelming number of non-catalytic residues in

enzymes compared to catalytic residues. This problem is tackled by: (1)

optimizing the classifier to maximize a performance criterion that considers both

Type I and Type II errors in the classification of catalytic and non-catalytic

residues; (2) under-sampling non-catalytic residues before SVM training; and (3)

during SVM training, penalizing errors in learning catalytic residues more than

errors in learning non-catalytic residues. Tested on four enzyme datasets, one

specifically designed by us to mimic the structural genomics scenario and three

previously evaluated datasets, our structure-based classifier is never inferior

to similar structure-based classifiers and comparable to classifiers that use

both structural and evolutionary features. In addition to the evaluation of the

performance of catalytic residue identification, we also present detailed case

studies on three proteins. This analysis suggests that many false positive

predictions may correspond to binding sites and other functional residues. A web

server that implements the method, our own-designed database, and the source code

of the programs are publicly available at

http://www.cs.bgu.ac.il/∼meshi/functionPrediction.

Copyright © 2011 Wiley-Liss, Inc.

DOI: 10.1002/prot.23020

PMCID: PMC3092797

PMID: 21491495 [Indexed for MEDLINE]

1707. BMC Immunol. 2011 May 27;12:33. doi: 10.1186/1471-2172-12-33.

LabKey Server NAb: a tool for analyzing, visualizing and sharing results from

neutralizing antibody assays.

Piehler B(1), Nelson EK, Eckels J, Ramsay S, Lum K, Wood B, Greene KM, Gao H,

Seaman MS, Montefiori DC, Igra M.

Author information:

(1)LabKey Software, Seattle, Washington 98102, USA.

BACKGROUND: Multiple types of assays allow sensitive detection of virus-specific

neutralizing antibodies. For example, the extent of antibody neutralization of

HIV-1, SIV and SHIV can be measured in the TZM-bl cell line through the degree of

luciferase reporter gene expression after infection. In the past, neutralization

curves and titers for this standard assay have been calculated using an Excel

macro. Updating all instances of such a macro with new techniques can be unwieldy

and introduce non-uniformity across multi-lab teams. Using Excel also poses

challenges in centrally storing, sharing and associating raw data files and

results.

RESULTS: We present LabKey Server's NAb tool for organizing, analyzing and

securely sharing data, files and results for neutralizing antibody (NAb) assays,

including the luciferase-based TZM-bl NAb assay. The customizable tool supports

high-throughput experiments and includes a graphical plate template designer,

allowing researchers to quickly adapt calculations to new plate layouts. The tool

calculates the percent neutralization for each serum dilution based on

luminescence measurements, fits a range of neutralization curves to titration

results and uses these curves to estimate the neutralizing antibody titers for

benchmark dilutions. Results, curve visualizations and raw data files are stored

in a database and shared through a secure, web-based interface. NAb results can

be integrated with other data sources based on sample identifiers. It is simple

to make results public after publication by updating folder security settings.

CONCLUSIONS: Standardized tools for analyzing, archiving and sharing assay

results can improve the reproducibility, comparability and reliability of results

obtained across many labs. LabKey Server and its NAb tool are freely available as

open source software at http://www.labkey.com under the Apache 2.0 license. Many

members of the HIV research community can also access the LabKey Server NAb tool

without installing the software by using the Atlas Science Portal

(https://atlas.scharp.org). Atlas is an installation of LabKey Server.

DOI: 10.1186/1471-2172-12-33

PMCID: PMC3115917

PMID: 21619655 [Indexed for MEDLINE]

1708. BMC Bioinformatics. 2011 May 26;12:207. doi: 10.1186/1471-2105-12-207.

HemeBIND: a novel method for heme binding residue prediction by combining

structural and sequence information.

Liu R(1), Hu J.

Author information:

(1)Department of Computer Science and Engineering, University of South Carolina,

Columbia, SC 29208, USA.

BACKGROUND: Accurate prediction of binding residues involved in the interactions

between proteins and small ligands is one of the major challenges in structural

bioinformatics. Heme is an essential and commonly used ligand that plays critical

roles in electron transfer, catalysis, signal transduction and gene expression.

Although much effort has been devoted to the development of various generic

algorithms for ligand binding site prediction over the last decade, no algorithm

has been specifically designed to complement experimental techniques for

identification of heme binding residues. Consequently, an urgent need is to

develop a computational method for recognizing these important residues.

RESULTS: Here we introduced an efficient algorithm HemeBIND for predicting heme

binding residues by integrating structural and sequence information. We

systematically investigated the characteristics of binding interfaces based on a

non-redundant dataset of heme-protein complexes. It was found that several

sequence and structural attributes such as evolutionary conservation, solvent

accessibility, depth and protrusion clearly illustrate the differences between

heme binding and non-binding residues. These features can then be separately used

or combined to build the structure-based classifiers using support vector machine

(SVM). The results showed that the information contained in these features is

largely complementary and their combination achieved the best performance. To

further improve the performance, an attempt has been made to develop a

post-processing procedure to reduce the number of false positives. In addition,

we built a sequence-based classifier based on SVM and sequence profile as an

alternative when only sequence information can be used. Finally, we employed a

voting method to combine the outputs of structure-based and sequence-based

classifiers, which demonstrated remarkably better performance than the individual

classifier alone.

CONCLUSIONS: HemeBIND is the first specialized algorithm used to predict binding

residues in protein structures for heme ligands. Extensive experiments indicated

that both the structure-based and sequence-based methods have effectively

identified heme binding residues while the complementary relationship between

them can result in a significant improvement in prediction performance. The value

of our method is highlighted through the development of HemeBIND web server that

is freely accessible at http://mleg.cse.sc.edu/hemeBIND/.

DOI: 10.1186/1471-2105-12-207

PMCID: PMC3124436

PMID: 21612668 [Indexed for MEDLINE]

1709. BMC Bioinformatics. 2011 May 25;12:199. doi: 10.1186/1471-2105-12-199.

MimoPro: a more efficient Web-based tool for epitope prediction using phage

display libraries.

Chen WH(1), Sun PP, Lu Y, Guo WW, Huang YX, Ma ZQ.

Author information:

(1)School of Computer Science and Information Technology, Northeast Normal

University, Changchun 130024, P.R. China.

BACKGROUND: A B-cell epitope is a group of residues on the surface of an antigen

which stimulates humoral responses. Locating these epitopes on antigens is

important for the purpose of effective vaccine design. In recent years, mapping

affinity-selected peptides screened from a random phage display library to the

native epitope has become popular in epitope prediction. These peptides, also

known as mimotopes, share the similar structure and function with the

corresponding native epitopes. Great effort has been made in using this

similarity between such mimotopes and native epitopes in prediction, which has

resulted in better outcomes than statistics-based methods can. However, it cannot

maintain a high degree of satisfaction in various circumstances.

RESULTS: In this study, we propose a new method that maps a group of mimotopes

back to a source antigen so as to locate the interacting epitope on the antigen.

The core of this method is a searching algorithm that is incorporated with both

dynamic programming (DP) and branch and bound (BB) optimization and operated on a

series of overlapping patches on the surface of a protein. These patches are then

transformed to a number of graphs using an adaptable distance threshold (ADT)

regulated by an appropriate compactness factor (CF), a novel parameter proposed

in this study. Compared with both Pep-3D-Search and PepSurf, two leading

graph-based search tools, on average from the results of 18 test cases, MimoPro,

the Web-based implementation of our proposed method, performed better in

sensitivity, precision, and Matthews correlation coefficient (MCC) than both did

in epitope prediction. In addition, MimoPro is significantly faster than both

Pep-3D-Search and PepSurf in processing.

CONCLUSIONS: Our search algorithm designed for processing well constructed graphs

using an ADT regulated by CF is more sensitive and significantly faster than

other graph-based approaches in epitope prediction. MimoPro is a viable

alternative to both PepSurf and Pep-3D-Search for epitope prediction in the same

kind, and freely accessible through the MimoPro server located at

http://informatics.nenu.edu.cn/MimoPro.

DOI: 10.1186/1471-2105-12-199

PMCID: PMC3124435

PMID: 21609501 [Indexed for MEDLINE]

1710. Bioinformatics. 2011 May 15;27(10):1447-8. doi: 10.1093/bioinformatics/btr156.

Epub 2011 Mar 30.

hmChIP: a database and web server for exploring publicly available human and

mouse ChIP-seq and ChIP-chip data.

Chen L(1), Wu G, Ji H.

Author information:

(1)Department of Biostatistics, The Johns Hopkins University Bloomberg School of

Public Health, Baltimore, MD 21205, USA.

hmChIP is a database of genome-wide chromatin immunoprecipitation (ChIP) data in

human and mouse. Currently, the database contains 2016 samples from 492 ChIP-seq

and ChIP-chip experiments, representing a total of 170 proteins and 11 069 914

protein-DNA interactions. A web server provides interface for database query.

Protein-DNA binding intensities can be retrieved from individual samples for

user-provided genomic regions. The retrieved intensities can be used to cluster

samples and genomic regions to facilitate exploration of combinatorial patterns,

cell-type dependencies, and cross-sample variability of protein-DNA

interactions.AVAILABILITY:

http://jilab.biostat.jhsph.edu/database/cgi-bin/hmChIP.pl.

DOI: 10.1093/bioinformatics/btr156

PMCID: PMC3087956

PMID: 21450710 [Indexed for MEDLINE]

1711. Bioinformatics. 2011 May 15;27(10):1438-9. doi: 10.1093/bioinformatics/btr096.

Epub 2011 Feb 23.

The MEMPACK alpha-helical transmembrane protein structure prediction server.

Nugent T(1), Ward S, Jones DT.

Author information:

(1)Department of Computer Science, University College London, London, UK.

t.nugent@cs.ucl.ac.uk

MOTIVATION: The experimental difficulties of alpha-helical transmembrane protein

structure determination make this class of protein an important target for

sequence-based structure prediction tools. The MEMPACK prediction server allows

users to submit a transmembrane protein sequence and returns transmembrane

topology, lipid exposure, residue contacts, helix-helix interactions and helical

packing arrangement predictions in both plain text and graphical formats using a

number of novel machine learning-based algorithms.

AVAILABILITY: The server can be accessed as a new component of the PSIPRED portal

by at http://bioinf.cs.ucl.ac.uk/psipred/.

DOI: 10.1093/bioinformatics/btr096

PMCID: PMC3087951

PMID: 21349872 [Indexed for MEDLINE]

1712. BMC Genomics. 2011 May 11;12:229. doi: 10.1186/1471-2164-12-229.

Meta-analysis of heterogeneous Down Syndrome data reveals consistent genome-wide

dosage effects related to neurological processes.

Vilardell M(1), Rasche A, Thormann A, Maschke-Dutz E, Pérez-Jurado LA, Lehrach H,

Herwig R.

Author information:

(1)Department of Vertebrate Genomics, Max-Planck-Institute for Molecular

Genetics, Berlin, Germany. nogales@molgen.mpg.de

BACKGROUND: Down syndrome (DS; trisomy 21) is the most common genetic cause of

mental retardation in the human population and key molecular networks

dysregulated in DS are still unknown. Many different experimental techniques have

been applied to analyse the effects of dosage imbalance at the molecular and

phenotypical level, however, currently no integrative approach exists that

attempts to extract the common information.

RESULTS: We have performed a statistical meta-analysis from 45 heterogeneous

publicly available DS data sets in order to identify consistent dosage effects

from these studies. We identified 324 genes with significant genome-wide dosage

effects, including well investigated genes like SOD1, APP, RUNX1 and DYRK1A as

well as a large proportion of novel genes (N = 62). Furthermore, we characterized

these genes using gene ontology, molecular interactions and promoter sequence

analysis. In order to judge relevance of the 324 genes for more general cerebral

pathologies we used independent publicly available microarry data from brain

studies not related with DS and identified a subset of 79 genes with potential

impact for neurocognitive processes. All results have been made available through

a web server under http://ds-geneminer.molgen.mpg.de/.

CONCLUSIONS: Our study represents a comprehensive integrative analysis of

heterogeneous data including genome-wide transcript levels in the domain of

trisomy 21. The detected dosage effects build a resource for further studies of

DS pathology and the development of new therapies.

DOI: 10.1186/1471-2164-12-229

PMCID: PMC3110572

PMID: 21569303 [Indexed for MEDLINE]

1713. J Theor Biol. 2011 May 7;276(1):229-49. doi: 10.1016/j.jtbi.2011.01.010. Epub

2011 Jan 26.

NL MIND-BEST: a web server for ligands and proteins

discovery--theoretic-experimental study of proteins of Giardia lamblia and new

compounds active against Plasmodium falciparum.

González-Díaz H(1), Prado-Prado F, Sobarzo-Sánchez E, Haddad M, Maurel Chevalley

S, Valentin A, Quetin-Leclercq J, Dea-Ayuela MA, Teresa Gomez-Muños M, Munteanu

CR, José Torres-Labandeira J, García-Mera X, Tapia RA, Ubeira FM.

Author information:

(1)Department of Microbiology and Parasitology, University of Santiago de

Compostela (USC), Galicia, Spain. gonzalezdiazh@yahoo.es

There are many protein ligands and/or drugs described with very different

affinity to a large number of target proteins or receptors. In this work, we

selected Ligands or Drug-target pairs (DTPs/nDTPs) of drugs with high

affinity/non-affinity for different targets. Quantitative Structure-Activity

Relationships (QSAR) models become a very useful tool in this context to

substantially reduce time and resources consuming experiments. Unfortunately most

QSAR models predict activity against only one protein target and/or have not been

implemented in the form of public web server freely accessible online to the

scientific community. To solve this problem, we developed here a multi-target

QSAR (mt-QSAR) classifier using the MARCH-INSIDE technique to calculate

structural parameters of drug and target plus one Artificial Neuronal Network

(ANN) to seek the model. The best ANN model found is a Multi-Layer Perceptron

(MLP) with profile MLP 20:20-15-1:1. This MLP classifies correctly 611 out of 678

DTPs (sensitivity=90.12%) and 3083 out of 3408 nDTPs (specificity=90.46%),

corresponding to training accuracy=90.41%. The validation of the model was

carried out by means of external predicting series. The model classifies

correctly 310 out of 338 DTPs (sensitivity=91.72%) and 1527 out of 1674 nDTP

(specificity=91.22%) in validation series, corresponding to total accuracy=91.30%

for validation series (predictability). This model favorably compares with other

ANN models developed in this work and Machine Learning classifiers published

before to address the same problem in different aspects. We implemented the

present model at web portal Bio-AIMS in the form of an online server called:

Non-Linear MARCH-INSIDE Nested Drug-Bank Exploration & Screening Tool (NL

MIND-BEST), which is located at URL:

http://miaja.tic.udc.es/Bio-AIMS/NL-MIND-BEST.php. This online tool is based on

PHP/HTML/Python and MARCH-INSIDE routines. Finally we illustrated two practical

uses of this server with two different experiments. In experiment 1, we report by

first time Quantum QSAR study, synthesis, characterization, and experimental

assay of antiplasmodial and cytotoxic activities of oxoisoaporphine alkaloids

derivatives as well as NL MIND-BEST prediction of potential target proteins. In

experiment 2, we report sampling, parasite culture, sample preparation, 2-DE,

MALDI-TOF, and -TOF/TOF MS, MASCOT search, MM/MD 3D structure modeling, and NL

MIND-BEST prediction for different peptides a new protein of the found in the

proteome of the human parasite Giardia lamblia, which is promising for

anti-parasite drug-targets discovery.

Copyright © 2011 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2011.01.010

PMID: 21277861 [Indexed for MEDLINE]

1714. Biochim Biophys Acta. 2011 May;1814(5):664-70. doi: 10.1016/j.bbapap.2011.03.004.

Epub 2011 Mar 21.

TMBHMM: a frequency profile based HMM for predicting the topology of

transmembrane beta barrel proteins and the exposure status of transmembrane

residues.

Singh NK(1), Goodman A, Walter P, Helms V, Hayat S.

Author information:

(1)Center for Bioinformatics, Saarland University, Saarbrücken, Germany.

nitesh\_summi@yahoo.com

Transmembrane beta barrel (TMB) proteins are found in the outer membranes of

bacteria, mitochondria and chloroplasts. TMBs are involved in a variety of

functions such as mediating flux of metabolites and active transport of

siderophores, enzymes and structural proteins, and in the translocation across or

insertion into membranes. We present here TMBHMM, a computational method based on

a hidden Markov model for predicting the structural topology of putative TMBs

from sequence. In addition to predicting transmembrane strands, TMBHMM also

predicts the exposure status (i.e., exposed to the membrane or hidden in the

protein structure) of the residues in the transmembrane region, which is a novel

feature of the TMBHMM method. Furthermore, TMBHMM can also predict the membrane

residues that are not part of beta barrel forming strands. The training of the

TMBHMM was performed on a non-redundant data set of 19 TMBs. The self-consistency

test yielded Q(2) accuracy of 0.87, Q(3) accuracy of 0.83, Matthews correlation

coefficient of 0.74 and SOV for beta strand of 0.95. In this self-consistency

test the method predicted 83% of transmembrane residues with correct exposure

status. On an unseen, non-redundant test data set of 10 proteins, the 2-state and

3-state TMBHMM prediction accuracies are around 73% and 72%, respectively, and

are comparable to other methods from the literature. The TMBHMM web server takes

an amino acid sequence or a multiple sequence alignment as an input and predicts

the exposure status and the structural topology as output. The TMBHMM web server

is available under the tmbhmm tab at:

http://service.bioinformatik.uni-saarland.de/tmx-site/.

Copyright © 2011 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.bbapap.2011.03.004

PMID: 21426944 [Indexed for MEDLINE]

1715. Bioinformatics. 2011 May 1;27(9):1322-3. doi: 10.1093/bioinformatics/btr119.

Rapid membrane protein topology prediction.

Hennerdal A(1), Elofsson A.

Author information:

(1)Department of Biochemistry and Biophysics, Stockholm Bioinformatics Center,

Center for Biomembrane Research, Swedish e-science Research Center, Stockholm

University, Stockholm, Sweden.

State-of-the-art methods for topology of α-helical membrane proteins are based on

the use of time-consuming multiple sequence alignments obtained from PSI-BLAST or

other sources. Here, we examine if it is possible to use the consensus of

topology prediction methods that are based on single sequences to obtain a

similar accuracy as the more accurate multiple sequence-based methods. Here, we

show that TOPCONS-single performs better than any of the other topology

prediction methods tested here, but ~6% worse than the best method that is

utilizing multiple sequence alignments.AVAILABILITY AND IMPLEMENTATION:

TOPCONS-single is available as a web server from http://single.topcons.net/ and

is also included for local installation from the web site. In addition,

consensus-based topology predictions for the entire international protein index

(IPI) is available from the web server and will be updated at regular intervals.

DOI: 10.1093/bioinformatics/btr119

PMCID: PMC3077071

PMID: 21493661 [Indexed for MEDLINE]

1716. Bioinformatics. 2011 May 1;27(9):1327-9. doi: 10.1093/bioinformatics/btr129. Epub

2011 Apr 1.

Rigid substructure search.

Shirvanyants D(1), Alexandrova AN, Dokholyan NV.

Author information:

(1)Department of Biochemistry and Biophysics, University of North Carolina at

Chapel Hill, NC 27599-7260, USA.

MOTIVATION: Identifying the location of binding sites on proteins is of

fundamental importance for a wide range of applications, including molecular

docking, de novo drug design, structure identification and comparison of

functional sites. Here we present Erebus, a web server that searches the entire

Protein Data Bank for a given substructure defined by a set of atoms of interest,

such as the binding scaffolds for small molecules. The identified substructure

contains atoms having the same names, belonging to same amino acids and separated

by the same distances (within a given tolerance) as the atoms of the query

structure. The accuracy of a match is measured by the root-mean-square deviation

or by the normal weight with a given variance. Tests show that our approach can

reliably locate rigid binding scaffolds of drugs and metal ions.

AVAILABILITY AND IMPLEMENTATION: We provide this service through a web server at

http://erebus.dokhlab.org.

DOI: 10.1093/bioinformatics/btr129

PMCID: PMC3138080

PMID: 21460026 [Indexed for MEDLINE]

1717. Bioinformatics. 2011 May 1;27(9):1224-30. doi: 10.1093/bioinformatics/btr108.

Epub 2011 Mar 2.

MemLoci: predicting subcellular localization of membrane proteins in eukaryotes.

Pierleoni A(1), Martelli PL, Casadio R.

Author information:

(1)Biocomputing Group, Computational Biology Network, via San Giacomo 9/2,

Bologna, Italy.

MOTIVATION: Subcellular localization is a key feature in the process of

functional annotation of both globular and membrane proteins. In the absence of

experimental data, protein localization is inferred on the basis of annotation

transfer upon sequence similarity search. However, predictive tools are necessary

when the localization of homologs is not known. This is so particularly for

membrane proteins. Furthermore, most of the available predictors of subcellular

localization are specifically trained on globular proteins and poorly perform on

membrane proteins.

RESULTS: Here we develop MemLoci, a new support vector machine-based tool that

discriminates three membrane protein localizations: plasma, internal and

organelle membrane. When tested on an independent set, MemLoci outperforms

existing methods, reaching an overall accuracy of 70% on predicting the location

in the three membrane types, with a generalized correlation coefficient as high

as 0.50.

AVAILABILITY: The MemLoci server is freely available on the web at:

http://mu2py.biocomp.unibo.it/memloci. Datasets described in the article can be

downloaded at the same site.

DOI: 10.1093/bioinformatics/btr108

PMID: 21367869 [Indexed for MEDLINE]

1718. Hum Mutat. 2011 May;32(5):557-63. doi: 10.1002/humu.21438. Epub 2011 Feb 22.

LOVD v.2.0: the next generation in gene variant databases.

Fokkema IF(1), Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT.

Author information:

(1)Center of Human and Clinical Genetics, Department of Human Genetics, Leiden

University Medical Center, Leiden, Nederland.

Locus-Specific DataBases (LSDBs) store information on gene sequence variation

associated with human phenotypes and are frequently used as a reference by

researchers and clinicians. We developed the Leiden Open-source Variation

Database (LOVD) as a platform-independent Web-based LSDB-in-a-Box package. LOVD

was designed to be easy to set up and maintain and follows the Human Genome

Variation Society (HGVS) recommendations. Here we describe LOVD v.2.0, which adds

enhanced flexibility and functionality and has the capacity to store sequence

variants in multiple genes per patient. To reduce redundancy, patient and

sequence variant data are stored in separate tables. Tables are linked to

generate connections between sequence variant data for each gene and every

patient. The dynamic structure allows database managers to add custom columns.

The database structure supports fast queries and allows storage of sequence

variants from high-throughput sequence analysis, as demonstrated by the

X-chromosomal Mental Retardation LOVD installation. LOVD contains measures to

ensure database security from unauthorized access. Currently, the LOVD Website

(http://www.LOVD.nl/) lists 71 public LOVD installations hosting 3,294 gene

variant databases with 199,000 variants in 84,000 patients. To promote LSDB

standardization and thereby database interoperability, we offer free server space

and help to establish an LSDB on our Leiden server.

© 2011 Wiley-Liss, Inc.

DOI: 10.1002/humu.21438

PMID: 21520333 [Indexed for MEDLINE]

1719. J Biomol NMR. 2011 May;50(1):43-57. doi: 10.1007/s10858-011-9478-4. Epub 2011 Mar

30.

SHIFTX2: significantly improved protein chemical shift prediction.

Han B(1), Liu Y, Ginzinger SW, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB, Canada.

A new computer program, called SHIFTX2, is described which is capable of rapidly

and accurately calculating diamagnetic (1)H, (13)C and (15)N chemical shifts from

protein coordinate data. Compared to its predecessor (SHIFTX) and to other

existing protein chemical shift prediction programs, SHIFTX2 is substantially

more accurate (up to 26% better by correlation coefficient with an RMS error that

is up to 3.3× smaller) than the next best performing program. It also provides

significantly more coverage (up to 10% more), is significantly faster (up to

8.5×) and capable of calculating a wider variety of backbone and side chain

chemical shifts (up to 6×) than many other shift predictors. In particular,

SHIFTX2 is able to attain correlation coefficients between experimentally

observed and predicted backbone chemical shifts of 0.9800 ((15)N), 0.9959

((13)Cα), 0.9992 ((13)Cβ), 0.9676 ((13)C'), 0.9714 ((1)HN), 0.9744 ((1)Hα) and

RMS errors of 1.1169, 0.4412, 0.5163, 0.5330, 0.1711, and 0.1231 ppm,

respectively. The correlation between SHIFTX2's predicted and observed side chain

chemical shifts is 0.9787 ((13)C) and 0.9482 ((1)H) with RMS errors of 0.9754 and

0.1723 ppm, respectively. SHIFTX2 is able to achieve such a high level of

accuracy by using a large, high quality database of training proteins (>190), by

utilizing advanced machine learning techniques, by incorporating many more

features (χ(2) and χ(3) angles, solvent accessibility, H-bond geometry, pH,

temperature), and by combining sequence-based with structure-based chemical shift

prediction techniques. With this substantial improvement in accuracy we believe

that SHIFTX2 will open the door to many long-anticipated applications of chemical

shift prediction to protein structure determination, refinement and validation.

SHIFTX2 is available both as a standalone program and as a web server (

http://www.shiftx2.ca ).

DOI: 10.1007/s10858-011-9478-4

PMCID: PMC3085061

PMID: 21448735 [Indexed for MEDLINE]

1720. J Comput Chem. 2011 May;32(7):1488-91. doi: 10.1002/jcc.21720. Epub 2011 Feb 1.

Web servers and services for electrostatics calculations with APBS and PDB2PQR.

Unni S, Huang Y, Hanson RM, Tobias M, Krishnan S, Li WW, Nielsen JE, Baker NA.

APBS and PDB2PQR are widely utilized free software packages for biomolecular

electrostatics calculations. Using the Opal toolkit, we have developed a Web

services framework for these software packages that enables the use of APBS and

PDB2PQR by users who do not have local access to the necessary amount of

computational capabilities. This not only increases accessibility of the software

to a wider range of scientists, educators, and students but also increases the

availability of electrostatics calculations on portable computing platforms.

Users can access this new functionality in two ways. First, an Opal-enabled

version of APBS is provided in current distributions, available freely on the

web. Second, we have extended the PDB2PQR web server to provide an interface for

the setup, execution, and visualization of electrostatic potentials as calculated

by APBS. This web interface also uses the Opal framework which ensures the

scalability needed to support the large APBS user community. Both of these

resources are available from the APBS/PDB2PQR website:

http://www.poissonboltzmann.org/.

Copyright © 2011 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.21720

PMCID: PMC3062090

PMID: 21425296 [Indexed for MEDLINE]

1721. Mol Divers. 2011 May;15(2):427-33. doi: 10.1007/s11030-010-9240-y. Epub 2010 Jul

22.

Prediction of mucin-type O-glycosylation sites by a two-staged strategy.

Cai Y(1), He J, Lu L.

Author information:

(1)Institute of System Biology, Shanghai University, 99 Shangda Road, 200244,

Shanghai, People's Republic of China. caiyudong@staff.shu.edu.cn

The mucin-type O-glycosylation of a protein is an important type of protein

post-translational modification. This process is mediated by a family of

UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases which transfer the

N-acetylgalactosamine (GalNAc) to the serine or threonine residues with unknown

specificity. In order to determine the glycosylation sites of a given protein, we

present a two-staged prediction method here, which first determines whether a

protein is a glycoprotein, and then determines the glycosylation sites of a

protein that has been predicted to be glycosylated in the first stage. In the

first stage, a protein is encoded by the protein families in PFAM, which is a

collective annotated database of classified protein families; then it is

predicted by a predictor trained by the training set. In the second stage,

nonapeptides of the predicted mucin-type glycoproteins, with serine or threonine

residues at their fifth sites, are represented by indices in AAIndex. Then, it is

predicted whether the nonapeptides are attached by GalNAc by a predictor, which

is constructed with features selected by feature selection methods [Maximum

Relevance Minimum Redundancy (mRMR) method and Incremental Feature Selection

method]. The prediction accuracy of the first stage is 94.9% validated by

Leave-One-Out validation method; the prediction accuracy of the second stage is

99.4%. These results show that this method is valuable to study the mucin-type

O-glycosylation. The analysis of the features used to construct the predictor of

the second stage confirms the previously obtained results from other groups. The

residues at position -1 and +3 have great impact on the prediction. Among other

amino acid indices, the indices about alpha and turn propensities and indices

about hydrophobicity of the residues in nonapeptide also influence the

recognition of the GalNAc transferases. A web server is available at

http://chemdata.shu.edu.cn/gal/.

DOI: 10.1007/s11030-010-9240-y

PMID: 20652405 [Indexed for MEDLINE]

1722. RNA. 2011 May;17(5):809-19. doi: 10.1261/rna.2474511. Epub 2011 Mar 25.

miTALOS: analyzing the tissue-specific regulation of signaling pathways by human

and mouse microRNAs.

Kowarsch A(1), Preusse M, Marr C, Theis FJ.

Author information:

(1)Institute for Bioinformatics and Systems Biology, Helmholtz Zentrum München,

Neuherberg, Germany. andreas.kowarsch@helmholtz-muenchen.de

Erratum in

RNA. 2012 May;18(5):1101.

MicroRNAs (miRNAs) are an important class of post-transcriptional regulators of

gene expression that are involved in various cellular and phenotypic processes. A

number of studies have shown that miRNA expression is induced by signaling

pathways. Moreover, miRNAs emerge as regulators of signaling pathways. Here, we

present the miTALOS web resource, which provides insight into miRNA-mediated

regulation of signaling pathways. As a novel feature, miTALOS considers the

tissue-specific expression signatures of miRNAs and target transcripts to improve

the analysis of miRNA regulation in biological pathways. MiTALOS identifies

potential pathway regulation by (i) an enrichment analysis of miRNA targets genes

and (ii) by using a proximity score to evaluate the functional role of miRNAs in

biological pathways by their network proximity. Moreover, miTALOS integrates five

different miRNA target prediction tools and two different signaling pathway

resources (KEGG and NCI). A graphical visualization of miRNA targets in both KEGG

and NCI PID signaling pathways is provided to illustrate their respective pathway

context. We perform a functional analysis on prostate cancer-related miRNAs and

are able to infer a model of miRNA-mediated regulation on tumor proliferation,

mobility and anti-apoptotic behavior. miTALOS provides novel features that

accomplish a substantial support to systematically infer regulation of signaling

pathways mediated by miRNAs. The web-server is freely accessible at

http://hmgu.de/cmb/mitalos.

DOI: 10.1261/rna.2474511

PMCID: PMC3078731

PMID: 21441347 [Indexed for MEDLINE]

1723. Physiol Genomics. 2011 Apr 27;43(8):457-60. doi:

10.1152/physiolgenomics.00178.2010. Epub 2011 Jan 18.

VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for

knowledge discovery.

Yang X(1), Ye Y, Wang G, Huang H, Yu D, Liang S.

Author information:

(1)Institute of Genetic Engineering, Southern Medical University, Guangzhou,

Guangdong Province, China.

In addition to many other genes, tissue-specific genes (TSGs) represent a set of

genes of great importance for human physiology. However, the links among TSGs,

diseases, and potential therapeutic agents are often missing, hidden, or too

scattered to find. There is a need to establish a knowledgebase for researchers

to share this and additional information in order to speed up discovery and

clinical practice. As an initiative toward systems biology, the VeryGene web

server was developed to fill this gap. A significant effort has been made to

integrate TSGs from two large-scale data analyses with respective information on

subcellular localization, Gene Ontology, Reactome, KEGG pathway, Mouse Genome

Informatics (MGI) Mammalian Phenotype, disease association, and targeting drugs.

The current release carefully selected 3,960 annotated TSGs derived from 127

normal human tissues and cell types, including 5,672 gene-disease and 2,171

drug-target relationships. In addition to being a specialized source for TSGs,

VeryGene can be used as a discovery tool by generating novel inferences. Some

inherently useful but hidden relations among genes, diseases, drugs, and other

important aspects can be inferred to form testable hypotheses. VeryGene is

available online at http://www.verygene.com.

DOI: 10.1152/physiolgenomics.00178.2010

PMID: 21245417 [Indexed for MEDLINE]

1724. Bioinformation. 2011 Apr 22;6(3):125-7.

PKSIIIexplorer: TSVM approach for predicting Type III polyketide synthase

proteins.

Vijayan M, Chandrika SK, Vasudevan SE.

PKSIIIexplorer, a web server based on 'transductive Support Vector Machine'

allows fast and reliable prediction of Type III polyketide synthase proteins. It

provides a simple unique platform to identify the probability of a particular

sequence, being a type III polyketide synthases or not with moderately high

accuracy. We hope that our method could serve as a useful program for the type

III polyketide researchers. The tool is available at

"http://type3pks.in/tsvm/pks3".ABBREVIATIONS: PKS - Polyketide synthase, CHS -

Chalcone synthase, SVM - Support vector machine, MCC - Matthews Correlation

Coefficient.

PMCID: PMC3089887

PMID: 21584189

1725. Bioinformatics. 2011 Apr 15;27(8):1185-6. doi: 10.1093/bioinformatics/btr097.

Epub 2011 Feb 23.

AGRA: analysis of gene ranking algorithms.

Kocbek S(1), Sætre R, Stiglic G, Kim JD, Pernek I, Tsuruoka Y, Kokol P, Ananiadou

S, Tsujii J.

Author information:

(1)Faculty of Health Sciences, University of Maribor, Maribor, Slovenia.

simon.kocbek@uni-mb.si

Often, the most informative genes have to be selected from different gene sets

and several computer gene ranking algorithms have been developed to cope with the

problem. To help researchers decide which algorithm to use, we developed the

analysis of gene ranking algorithms (AGRA) system that offers a novel technique

for comparing ranked lists of genes. The most important feature of AGRA is that

no previous knowledge of gene ranking algorithms is needed for their comparison.

Using the text mining system finding-associated concepts with text analysis. AGRA

defines what we call biomedical concept space (BCS) for each gene list and offers

a comparison of the gene lists in six different BCS categories. The uploaded gene

lists can be compared using two different methods. In the first method, the

overlap between each pair of two gene lists of BCSs is calculated. The second

method offers a text field where a specific biomedical concept can be entered.

AGRA searches for this concept in each gene lists' BCS, highlights the rank of

the concept and offers a visual representation of concepts ranked above and below

it.AVAILABILITY AND IMPLEMENTATION: Available at http://agra.fzv.uni-mb.si/,

implemented in Java and running on the Glassfish server.

CONTACT: simon.kocbek@uni-mb.si.

DOI: 10.1093/bioinformatics/btr097

PMCID: PMC3072556

PMID: 21349873 [Indexed for MEDLINE]

1726. Bioinformatics. 2011 Apr 15;27(8):1174-5. doi: 10.1093/bioinformatics/btr086.

Epub 2011 Feb 23.

BiC: a web server for calculating bimodality of coexpression between gene and

protein networks.

Linderman GC(1), Patel VN, Chance MR, Bebek G.

Author information:

(1)Department of Biomedical Engineering, Case Center for Proteomics and

Bioinformatics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland,

OH 44106, USA.

Bimodal patterns of expression have recently been shown to be useful not only in

prioritizing genes that distinguish phenotypes, but also in prioritizing network

models that correlate with proteomic evidence. In particular, subgroups of

strongly coexpressed gene pairs result in an increased variance of the

correlation distribution. This variance, a measure of association between sets of

genes (or proteins), can be summarized as the bimodality of coexpression (BiC).

We developed an online tool to calculate the BiC for user-defined gene lists and

associated mRNA expression data. BiC is a comprehensive application that provides

researchers with the ability to analyze both publicly available and

user-collected array data.AVAILABILITY: The freely available web service and the

documentation can be accessed at http://gurkan.case.edu/software.

CONTACT: gurkan@case.edu.

DOI: 10.1093/bioinformatics/btr086

PMCID: PMC3072551

PMID: 21345871 [Indexed for MEDLINE]

1727. PLoS One. 2011 Apr 13;6(4):e18476. doi: 10.1371/journal.pone.0018476.

Prediction of antimicrobial peptides based on sequence alignment and feature

selection methods.

Wang P(1), Hu L, Liu G, Jiang N, Chen X, Xu J, Zheng W, Li L, Tan M, Chen Z, Song

H, Cai YD, Chou KC.

Author information:

(1)Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences,

Tianjin, China.

Antimicrobial peptides (AMPs) represent a class of natural peptides that form a

part of the innate immune system, and this kind of 'nature's antibiotics' is

quite promising for solving the problem of increasing antibiotic resistance. In

view of this, it is highly desired to develop an effective computational method

for accurately predicting novel AMPs because it can provide us with more

candidates and useful insights for drug design. In this study, a new method for

predicting AMPs was implemented by integrating the sequence alignment method and

the feature selection method. It was observed that, the overall jackknife success

rate by the new predictor on a newly constructed benchmark dataset was over

80.23%, and the Mathews correlation coefficient is 0.73, indicating a good

prediction. Moreover, it is indicated by an in-depth feature analysis that the

results are quite consistent with the previously known knowledge that some amino

acids are preferential in AMPs and that these amino acids do play an important

role for the antimicrobial activity. For the convenience of most experimental

scientists who want to use the prediction method without the interest to follow

the mathematical details, a user-friendly web-server is provided at

http://amp.biosino.org/.

DOI: 10.1371/journal.pone.0018476

PMCID: PMC3076375

PMID: 21533231 [Indexed for MEDLINE]

1728. BMC Res Notes. 2011 Apr 12;4:121. doi: 10.1186/1756-0500-4-121.

mmView: a web-based viewer of the mmCIF format.

Cech P(1), Svozil D.

Author information:

(1)Laboratory of Informatics and Chemistry, Faculty of Chemical Technology,

Institute of Chemical Technology, Technická 5, CZ-166 28 Prague, Czech Republic.

daniel.svozil@vscht.cz.

BACKGROUND: Structural biomolecular data are commonly stored in the PDB format.

The PDB format is widely supported by software vendors because of its simplicity

and readability. However, the PDB format cannot fully address many informatics

challenges related to the growing amount of structural data. To overcome the

limitations of the PDB format, a new textual format mmCIF was released in June

1997 in its version 1.0. mmCIF provides extra information which has the advantage

of being in a computer readable form. However, this advantage becomes a

disadvantage if a human must read and understand the stored data. While software

tools exist to help to prepare mmCIF files, the number of available systems

simplifying the comprehension and interpretation of the mmCIF files is limited.

FINDINGS: In this paper we present mmView - a cross-platform web-based

application that allows to explore comfortably the structural data of

biomacromolecules stored in the mmCIF format. The mmCIF categories can be easily

browsed in a tree-like structure, and the corresponding data are presented in a

well arranged tabular form. The application also allows to display and

investigate biomolecular structures via an integrated Java application Jmol.

CONCLUSIONS: The mmView software system is primarily intended for educational

purposes, but it can also serve as a useful research tool. The mmView application

is offered in two flavors: as an open-source stand-alone application (available

from http://sourceforge.net/projects/mmview) that can be installed on the user's

computer, and as a publicly available web server.

DOI: 10.1186/1756-0500-4-121

PMCID: PMC3094370

PMID: 21486459

1729. Bioinformatics. 2011 Apr 1;27(7):1017-8. doi: 10.1093/bioinformatics/btr064. Epub

2011 Feb 16.

FIMO: scanning for occurrences of a given motif.

Grant CE(1), Bailey TL, Noble WS.

Author information:

(1)Department of Genome Sciences, University of Washington, Seattle, WA, USA.

A motif is a short DNA or protein sequence that contributes to the biological

function of the sequence in which it resides. Over the past several decades, many

computational methods have been described for identifying, characterizing and

searching with sequence motifs. Critical to nearly any motif-based sequence

analysis pipeline is the ability to scan a sequence database for occurrences of a

given motif described by a position-specific frequency matrix.RESULTS: We

describe Find Individual Motif Occurrences (FIMO), a software tool for scanning

DNA or protein sequences with motifs described as position-specific scoring

matrices. The program computes a log-likelihood ratio score for each position in

a given sequence database, uses established dynamic programming methods to

convert this score to a P-value and then applies false discovery rate analysis to

estimate a q-value for each position in the given sequence. FIMO provides output

in a variety of formats, including HTML, XML and several Santa Cruz Genome

Browser formats. The program is efficient, allowing for the scanning of DNA

sequences at a rate of 3.5 Mb/s on a single CPU.

AVAILABILITY AND IMPLEMENTATION: FIMO is part of the MEME Suite software toolkit.

A web server and source code are available at http://meme.sdsc.edu.

DOI: 10.1093/bioinformatics/btr064

PMCID: PMC3065696

PMID: 21330290 [Indexed for MEDLINE]

1730. Bioinformatics. 2011 Apr 1;27(7):968-72. doi: 10.1093/bioinformatics/btr061. Epub

2011 Feb 7.

HLA\*IMP--an integrated framework for imputing classical HLA alleles from SNP

genotypes.

Dilthey AT(1), Moutsianas L, Leslie S, McVean G.

Author information:

(1)Department of Statistics, University of Oxford, Oxford, UK.

MOTIVATION: Genetic variation at classical HLA alleles influences many

phenotypes, including susceptibility to autoimmune disease, resistance to

pathogens and the risk of adverse drug reactions. However, classical HLA typing

methods are often prohibitively expensive for large-scale studies. We previously

described a method for imputing classical alleles from linked SNP genotype data.

Here, we present a modification of the original algorithm implemented in a freely

available software suite that combines local data preparation and QC with

probabilistic imputation through a remote server.

RESULTS: We introduce two modifications to the original algorithm. First, we

present a novel SNP selection function that leads to pronounced increases (up by

40% in some scenarios) in call rate. Second, we develop a parallelized model

building algorithm that allows us to process a reference set of over 2500

individuals. In a validation experiment, we show that our framework produces

highly accurate HLA type imputations at class I and class II loci for independent

datasets: at call rates of 95-99%, imputation accuracy is between 92% and 98% at

the four-digit level and over 97% at the two-digit level. We demonstrate utility

of the method through analysis of a genome-wide association study for psoriasis

where there is a known classical HLA risk allele (HLA-C\*06:02). We show that the

imputed allele shows stronger association with disease than any single SNP within

the region. The imputation framework, HLA\*IMP, provides a powerful tool for

dissecting the architecture of genetic risk within the HLA.

AVAILABILITY: HLA\*IMP, implemented in C++ and Perl, is available from

http://oxfordhla.well.ox.ac.uk and is free for academic use.

DOI: 10.1093/bioinformatics/btr061

PMCID: PMC3065693

PMID: 21300701 [Indexed for MEDLINE]

1731. Bioinformatics. 2011 Apr 1;27(7):925-32. doi: 10.1093/bioinformatics/btr044. Epub

2011 Feb 3.

GOSSIP: a method for fast and accurate global alignment of protein structures.

Kifer I(1), Nussinov R, Wolfson HJ.

Author information:

(1)School of Computer Science, Raymond and Beverly Sackler Faculty of Exact

Sciences, Sackler Institute of Molecular Medicine, Tel Aviv University, Tel Aviv,

Israel.

MOTIVATION: The database of known protein structures (PDB) is increasing rapidly.

This results in a growing need for methods that can cope with the vast amount of

structural data. To analyze the accumulating data, it is important to have a fast

tool for identifying similar structures and clustering them by structural

resemblance. Several excellent tools have been developed for the comparison of

protein structures. These usually address the task of local structure alignment,

an important yet computationally intensive problem due to its complexity. It is

difficult to use such tools for comparing a large number of structures to each

other at a reasonable time.

RESULTS: Here we present GOSSIP, a novel method for a global all-against-all

alignment of any set of protein structures. The method detects similarities

between structures down to a certain cutoff (a parameter of the program), hence

allowing it to detect similar structures at a much higher speed than local

structure alignment methods. GOSSIP compares many structures in times which are

several orders of magnitude faster than well-known available structure alignment

servers, and it is also faster than a database scanning method. We evaluate

GOSSIP both on a dataset of short structural fragments and on two large

sequence-diverse structural benchmarks. Our conclusions are that for a threshold

of 0.6 and above, the speed of GOSSIP is obtained with no compromise of the

accuracy of the alignments or of the number of detected global similarities.

AVAILABILITY: A server, as well as an executable for download, are available at

http://bioinfo3d.cs.tau.ac.il/gossip/.

DOI: 10.1093/bioinformatics/btr044

PMCID: PMC3065682

PMID: 21296751 [Indexed for MEDLINE]

1732. Bioinformatics. 2011 Apr 1;27(7):903-11. doi: 10.1093/bioinformatics/btr040. Epub

2011 Jan 28.

Conveyor: a workflow engine for bioinformatic analyses.

Linke B(1), Giegerich R, Goesmann A.

Author information:

(1)Bioinformatics Resource Faciliy, Center for Biotechnology and Faculty of

Technology, Bielefeld University, 33615 Bielefeld, Germany.

blinke@ceBiTec.Uni-Bielefeld.De

MOTIVATION: The rapidly increasing amounts of data available from new

high-throughput methods have made data processing without automated pipelines

infeasible. As was pointed out in several publications, integration of data and

analytic resources into workflow systems provides a solution to this problem,

simplifying the task of data analysis. Various applications for defining and

running workflows in the field of bioinformatics have been proposed and

published, e.g. Galaxy, Mobyle, Taverna, Pegasus or Kepler. One of the main aims

of such workflow systems is to enable scientists to focus on analysing their

datasets instead of taking care for data management, job management or monitoring

the execution of computational tasks. The currently available workflow systems

achieve this goal, but fundamentally differ in their way of executing workflows.

RESULTS: We have developed the Conveyor software library, a multitiered generic

workflow engine for composition, execution and monitoring of complex workflows.

It features an open, extensible system architecture and concurrent program

execution to exploit resources available on modern multicore CPU hardware. It

offers the ability to build complex workflows with branches, loops and other

control structures. Two example use cases illustrate the application of the

versatile Conveyor engine to common bioinformatics problems.

AVAILABILITY: The Conveyor application including client and server are available

at http://conveyor.cebitec.uni-bielefeld.de.

DOI: 10.1093/bioinformatics/btr040

PMID: 21278189 [Indexed for MEDLINE]

1733. Indian J Biochem Biophys. 2011 Apr;48(2):106-10.

MAPS: an interactive web server for membrane annotation of transmembrane protein

structures.

Cheema J(1), Basu G.

Author information:

(1)Department of Biophysics, Bose Institute, P-1/12 CIT Scheme VIIM, Kolkata 700

054, India. jitender.cheema@bbsrc.ac.uk

The exact positioning of the membrane in transmembrane (TM) proteins plays

important functional roles. Yet, the structures of TM proteins in protein data

bank (pdb) have no information about the explicit position of the membrane. Using

a simple hydrophobic lipid-protein mismatch energy function and a flexible

lipid/water boundary, the position of lipid bilayer for representative TM

proteins in pdb have been annotated. A web server called MAPS (Membrane

Annotation of Protein Structures; available at:

http://www.boseinst.ernet.in/gautam/maps) has been set up that allows the user to

interactively analyze membrane-protein orientations of any uploaded pdb structure

with user-defined membrane flexibility parameters.

PMID: 21682142 [Indexed for MEDLINE]

1734. J Proteome Res. 2011 Apr 1;10(4):1794-805. doi: 10.1021/pr101065j. Epub 2011 Feb

22.

Andromeda: a peptide search engine integrated into the MaxQuant environment.

Cox J(1), Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M.

Author information:

(1)Department of Proteomics and Signal Transduction, Max-Planck-Institute of

Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. cox@biochem.mpg.de

A key step in mass spectrometry (MS)-based proteomics is the identification of

peptides in sequence databases by their fragmentation spectra. Here we describe

Andromeda, a novel peptide search engine using a probabilistic scoring model. On

proteome data, Andromeda performs as well as Mascot, a widely used commercial

search engine, as judged by sensitivity and specificity analysis based on target

decoy searches. Furthermore, it can handle data with arbitrarily high fragment

mass accuracy, is able to assign and score complex patterns of post-translational

modifications, such as highly phosphorylated peptides, and accommodates extremely

large databases. The algorithms of Andromeda are provided. Andromeda can function

independently or as an integrated search engine of the widely used MaxQuant

computational proteomics platform and both are freely available at

www.maxquant.org. The combination enables analysis of large data sets in a simple

analysis workflow on a desktop computer. For searching individual spectra

Andromeda is also accessible via a web server. We demonstrate the flexibility of

the system by implementing the capability to identify cofragmented peptides,

significantly improving the total number of identified peptides.

DOI: 10.1021/pr101065j

PMID: 21254760 [Indexed for MEDLINE]

1735. J Proteome Res. 2011 Apr 1;10(4):1698-718. doi: 10.1021/pr101009e. Epub 2011 Feb

24.

MIND-BEST: Web server for drugs and target discovery; design, synthesis, and

assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein

from Trichomonas gallinae.

González-Díaz H(1), Prado-Prado F, García-Mera X, Alonso N, Abeijón P, Caamaño O,

Yáñez M, Munteanu CR, Pazos A, Dea-Ayuela MA, Gómez-Muñoz MT, Garijo MM, Sansano

J, Ubeira FM.

Author information:

(1)Department of Microbiology and Parasitology, University of Santiago de

Compostela, Spain. gonzalezdiazh@yahoo.es

Many drugs with very different affinity to a large number of receptors are

described. Thus, in this work, we selected drug-target pairs (DTPs/nDTPs) of

drugs with high affinity/nonaffinity for different targets. Quantitative

structure-activity relationship (QSAR) models become a very useful tool in this

context because they substantially reduce time and resource-consuming

experiments. Unfortunately, most QSAR models predict activity against only one

protein target and/or they have not been implemented on a public Web server yet,

freely available online to the scientific community. To solve this problem, we

developed a multitarget QSAR (mt-QSAR) classifier combining the MARCH-INSIDE

software for the calculation of the structural parameters of drug and target with

the linear discriminant analysis (LDA) method in order to seek the best model.

The accuracy of the best LDA model was 94.4% (3,859/4,086 cases) for training and

94.9% (1,909/2,012 cases) for the external validation series. In addition, we

implemented the model into the Web portal Bio-AIMS as an online server entitled

MARCH-INSIDE Nested Drug-Bank Exploration & Screening Tool (MIND-BEST), located

at http://miaja.tic.udc.es/Bio-AIMS/MIND-BEST.php . This online tool is based on

PHP/HTML/Python and MARCH-INSIDE routines. Finally, we illustrated two practical

uses of this server with two different experiments. In experiment 1, we report

for the first time a MIND-BEST prediction, synthesis, characterization, and MAO-A

and MAO-B pharmacological assay of eight rasagiline derivatives, promising for

anti-Parkinson drug design. In experiment 2, we report sampling, parasite

culture, sample preparation, 2-DE, MALDI-TOF and -TOF/TOF MS, MASCOT search, 3D

structure modeling with LOMETS, and MIND-BEST prediction for different peptides

as new protein of the found in the proteome of the bird parasite Trichomonas

gallinae, which is promising for antiparasite drug targets discovery.

DOI: 10.1021/pr101009e

PMID: 21184613 [Indexed for MEDLINE]

1736. Nan Fang Yi Ke Da Xue Xue Bao. 2011 Apr;31(4):610-4.

[Design and realization of a microarray data analysis platform].

[Article in Chinese]

Sun XH(1), Guo YB, Liu N, Ma L, Deng QK.

Author information:

(1)Institude of Bioinformatics, School of Preclinical Medicine, Southern Medical

University, Guangzhou 510515, China. heather\_sun@126.com

OBJECTIVE: To design a platform for microarray data analysis and processing in

the browser/server mode running in Linux operating system.

METHODS: Based on the Apache HTTP server, the platform, programmed with Perl

language, integrated R language and Bioconductor packages for processing and

analysis of the input data of oligonucleotide arrays and two-color spotted

arrays. Users were allowed to submit data and parameter configurations to the

platform via the web page, and the results of analysis were also returned via the

web page.

RESULTS: With an easy operation and high performance, the platform fulfilled the

functions of processing, quality assessment, biological annotation and

statistical analysis of the data from oligonucleotide arrays and two-color

spotted arrays. Using the platform, we analyzed the gene expression profiles in

Mtb-stimulated macrophages of three clinical phenotypes, namely latent TB (LTB),

pulmonary (PTB) and meningeal (TBM), and obtained valuable clues for identifying

tuberculosis susceptibility genes. We also analyzed the effect of INH treatment

on Mycobacterium tuberculosis gene expression in various dormancy models, such as

hypoxia and KatG mutant, and found that a set of genes responded to INH treatment

during exponential growth but not in dormancy, and that the overall number of

differentially regulated genes was reduced in the cells in low metabolic state.

CONCLUSION: The platform we have constructed integrates comprehensive resources,

and with a user-friendly interface, allows direct result visualization to

facilitate microarray data analysis.

PMID: 21515453 [Indexed for MEDLINE]

1737. Proteins. 2011 Apr;79(4):1230-9. doi: 10.1002/prot.22958. Epub 2011 Jan 25.

Prediction of RNA-binding residues in proteins from primary sequence using an

enriched random forest model with a novel hybrid feature.

Ma X(1), Guo J, Wu J, Liu H, Yu J, Xie J, Sun X.

Author information:

(1)State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096,

People's Republic of China.

The identification of RNA-binding residues in proteins is important in several

areas such as protein function, posttranscriptional regulation and drug design.

We have developed PRBR (Prediction of RNA Binding Residues), a novel method for

identifying RNA-binding residues from amino acid sequences. Our method combines a

hybrid feature with the enriched random forest (ERF) algorithm. The hybrid

feature is composed of predicted secondary structure information and three novel

features: evolutionary information combined with conservation information of the

physicochemical properties of amino acids and the information about dependency of

amino acids with regards to polarity-charge and hydrophobicity in the protein

sequences. Our results demonstrate that the PRBR model achieves 0.5637 Matthew's

correlation coefficient (MCC) and 88.63% overall accuracy (ACC) with 53.70%

sensitivity (SE) and 96.97% specificity (SP). By comparing the performance of

each feature we found that all three novel features contribute to the improved

predictions. Area under the curve (AUC) statistics from receiver operating

characteristic curve analysis was compared between PRBR model and other models.

The results show that PRBR achieves the highest AUC value (0.8675) which

represents that PRBR attains excellent performance on predicting the RNA-binding

residues in proteins. The PRBR web-server implementation is freely available at

http://www.cbi.seu.edu.cn/PRBR/.

Copyright © 2011 Wiley-Liss, Inc.

DOI: 10.1002/prot.22958

PMID: 21268114 [Indexed for MEDLINE]

1738. BMC Bioinformatics. 2011 Mar 31;12:90. doi: 10.1186/1471-2105-12-90.

Using context to improve protein domain identification.

Ochoa A(1), Llinás M, Singh M.

Author information:

(1)Department of Molecular Biology, Princeton University, Princeton, NJ, USA.

BACKGROUND: Identifying domains in protein sequences is an important step in

protein structural and functional annotation. Existing domain recognition methods

typically evaluate each domain prediction independently of the rest. However, the

majority of proteins are multidomain, and pairwise domain co-occurrences are

highly specific and non-transitive.

RESULTS: Here, we demonstrate how to exploit domain co-occurrence to boost weak

domain predictions that appear in previously observed combinations, while

penalizing higher confidence domains if such combinations have never been

observed. Our framework, Domain Prediction Using Context (dPUC), incorporates

pairwise "context" scores between domains, along with traditional domain scores

and thresholds, and improves domain prediction across a variety of organisms from

bacteria to protozoa and metazoa. Among the genomes we tested, dPUC is most

successful at improving predictions for the poorly-annotated malaria parasite

Plasmodium falciparum, for which over 38% of the genome is currently unannotated.

Our approach enables high-confidence annotations in this organism and the

identification of orthologs to many core machinery proteins conserved in all

eukaryotes, including those involved in ribosomal assembly and other RNA

processing events, which surprisingly had not been previously known.

CONCLUSIONS: Overall, our results demonstrate that this new context-based

approach will provide significant improvements in domain and function prediction,

especially for poorly understood genomes for which the need for additional

annotations is greatest. Source code for the algorithm is available under a GPL

open source license at http://compbio.cs.princeton.edu/dpuc/. Pre-computed

results for our test organisms and a web server are also available at that

location.

DOI: 10.1186/1471-2105-12-90

PMCID: PMC3090354

PMID: 21453511 [Indexed for MEDLINE]

1739. PLoS One. 2011 Mar 30;6(3):e18258. doi: 10.1371/journal.pone.0018258.

iLoc-Euk: a multi-label classifier for predicting the subcellular localization of

singleplex and multiplex eukaryotic proteins.

Chou KC(1), Wu ZC, Xiao X.

Author information:

(1)Gordon Life Science Institute, San Diego, California, United States of

America. kcchou@gordonlifescience.org

Predicting protein subcellular localization is an important and difficult

problem, particularly when query proteins may have the multiplex character, i.e.,

simultaneously exist at, or move between, two or more different subcellular

location sites. Most of the existing protein subcellular location predictor can

only be used to deal with the single-location or "singleplex" proteins. Actually,

multiple-location or "multiplex" proteins should not be ignored because they

usually posses some unique biological functions worthy of our special notice. By

introducing the "multi-labeled learning" and "accumulation-layer scale", a new

predictor, called iLoc-Euk, has been developed that can be used to deal with the

systems containing both singleplex and multiplex proteins. As a demonstration,

the jackknife cross-validation was performed with iLoc-Euk on a benchmark dataset

of eukaryotic proteins classified into the following 22 location sites: (1)

acrosome, (2) cell membrane, (3) cell wall, (4) centriole, (5) chloroplast, (6)

cyanelle, (7) cytoplasm, (8) cytoskeleton, (9) endoplasmic reticulum, (10)

endosome, (11) extracellular, (12) Golgi apparatus, (13) hydrogenosome, (14)

lysosome, (15) melanosome, (16) microsome (17) mitochondrion, (18) nucleus, (19)

peroxisome, (20) spindle pole body, (21) synapse, and (22) vacuole, where none of

proteins included has ≥25% pairwise sequence identity to any other in a same

subset. The overall success rate thus obtained by iLoc-Euk was 79%, which is

significantly higher than that by any of the existing predictors that also have

the capacity to deal with such a complicated and stringent system. As a

user-friendly web-server, iLoc-Euk is freely accessible to the public at the

web-site http://icpr.jci.edu.cn/bioinfo/iLoc-Euk. It is anticipated that iLoc-Euk

may become a useful bioinformatics tool for Molecular Cell Biology, Proteomics,

System Biology, and Drug Development Also, its novel approach will further

stimulate the development of predicting other protein attributes.

DOI: 10.1371/journal.pone.0018258

PMCID: PMC3068162

PMID: 21483473 [Indexed for MEDLINE]

1740. J Chem Inf Model. 2011 Mar 28;51(3):624-34. doi: 10.1021/ci1003174. Epub 2011 Feb

28.

ReverseScreen3D: a structure-based ligand matching method to identify protein

targets.

Kinnings SL(1), Jackson RM.

Author information:

(1)Institute of Molecular and Cellular Biology, Faculty of Biological Sciences,

University of Leeds, Leeds, United Kingdom.

Ligand promiscuity, which is now recognized as an extremely common phenomenon, is

a major underlying cause of drug toxicity. We have developed a new reverse

virtual screening (VS) method called ReverseScreen3D, which can be used to

predict the potential protein targets of a query compound of interest. The method

uses a 2D fingerprint-based method to select a ligand template from each unique

binding site of each protein within a target database. The target database

contains only the structurally determined bioactive conformations of known

ligands. The 2D comparison is followed by a 3D structural comparison to the

selected query ligand using a geometric matching method, in order to prioritize

each target binding site in the database. We have evaluated the performance of

the ReverseScreen2D and 3D methods using a diverse set of small molecule protein

inhibitors known to have multiple targets, and have shown that they are able to

provide a highly significant enrichment of true targets in the database.

Furthermore, we have shown that the 3D structural comparison improves early

enrichment when compared with the 2D method alone, and that the 3D method

performs well even in the absence of 2D similarity to the template ligands. By

carrying out further experimental screening on the prioritized list of targets,

it may be possible to determine the potential targets of a new compound or

determine the off-targets of an existing drug. The ReverseScreen3D method has

been incorporated into a Web server, which is freely available at

http://www.modelling.leeds.ac.uk/ReverseScreen3D .

DOI: 10.1021/ci1003174

PMID: 21361385 [Indexed for MEDLINE]

1741. Front Genet. 2011 Mar 25;2:14. doi: 10.3389/fgene.2011.00014. eCollection 2011.

GRFT - Genetic Records Family Tree Web Applet.

Pimentel S(1), Walbot V, Fernandes J.

Author information:

(1)Department of Biology, Stanford University Stanford, CA, USA.

Current software for storing and displaying records of genetic crosses does not

provide an easy way to determine the lineage of an individual. The genetic

records family tree (GRFT) applet processes records of genetic crosses and allows

researchers to quickly visualize lineages using a family tree construct and to

access other information from these records using any Internet browser. Users

select from three display features: (1) a family tree view which displays a

color-coded family tree for an individual, (2) a sequential list of crosses, and

(3) a list of crosses matching user-defined search criteria. Each feature

contains options to specify the number of records shown and the latter two

contain an option to filter results by the owner of the cross. The family tree

feature is interactive, displaying a popup box with genetic information when the

user mouses over an individual and allowing the user to draw a new tree by

clicking on any individual in the current tree. The applet is written in

JavaScript and reads genetic records from a tab-delimited text file on the

server, so it is cross-platform, can be accessed by anyone with an Internet

connection, and supports almost instantaneous generation of new trees and table

lists. Researchers can use the tool with their own genetic cross records for any

sexually reproducing organism. No additional software is required and with only

minor modifications to the script, researchers can add their own custom columns.

GRFT's speed, versatility, and low overhead make it an effective and innovative

visualization method for genetic records. A sample tool is available at

http://stanford.edu/walbot/grft-sample.html.

DOI: 10.3389/fgene.2011.00014

PMCID: PMC3270322

PMID: 22303311

1742. PLoS One. 2011 Mar 25;6(3):e17695. doi: 10.1371/journal.pone.0017695.

CPORT: a consensus interface predictor and its performance in prediction-driven

docking with HADDOCK.

de Vries SJ(1), Bonvin AM.

Author information:

(1)Faculty of Science, Bijvoet Center for Biomolecular Research, Utrecht

University, Utrecht, The Netherlands. a.m.j.j.bonvin@uu.nl

BACKGROUND: Macromolecular complexes are the molecular machines of the cell.

Knowledge at the atomic level is essential to understand and influence their

function. However, their number is huge and a significant fraction is extremely

difficult to study using classical structural methods such as NMR and X-ray

crystallography. Therefore, the importance of large-scale computational

approaches in structural biology is evident. This study combines two of these

computational approaches, interface prediction and docking, to obtain

atomic-level structures of protein-protein complexes, starting from their unbound

components.

METHODOLOGY/PRINCIPAL FINDINGS: Here we combine six interface prediction web

servers into a consensus method called CPORT (Consensus Prediction Of interface

Residues in Transient complexes). We show that CPORT gives more stable and

reliable predictions than each of the individual predictors on its own. A

protocol was developed to integrate CPORT predictions into our data-driven

docking program HADDOCK. For cases where experimental information is limited,

this prediction-driven docking protocol presents an alternative to ab initio

docking, the docking of complexes without the use of any information.

Prediction-driven docking was performed on a large and diverse set of

protein-protein complexes in a blind manner. Our results indicate that the

performance of the HADDOCK-CPORT combination is competitive with ZDOCK-ZRANK, a

state-of-the-art ab initio docking/scoring combination. Finally, the original

interface predictions could be further improved by interface post-prediction

(contact analysis of the docking solutions).

CONCLUSIONS/SIGNIFICANCE: The current study shows that blind, prediction-driven

docking using CPORT and HADDOCK is competitive with ab initio docking methods.

This is encouraging since prediction-driven docking represents the absolute

bottom line for data-driven docking: any additional biological knowledge will

greatly improve the results obtained by prediction-driven docking alone. Finally,

the fact that original interface predictions could be further improved by

interface post-prediction suggests that prediction-driven docking has not yet

been pushed to the limit. A web server for CPORT is freely available at

http://haddock.chem.uu.nl/services/CPORT.

DOI: 10.1371/journal.pone.0017695

PMCID: PMC3064578

PMID: 21464987 [Indexed for MEDLINE]

1743. PLoS One. 2011 Mar 21;6(3):e17911. doi: 10.1371/journal.pone.0017911.

GOBO: gene expression-based outcome for breast cancer online.

Ringnér M(1), Fredlund E, Häkkinen J, Borg Å, Staaf J.

Author information:

(1)Department of Oncology, Clinical Sciences, Lund University and Skåne

University Hospital, Lund, Sweden.

Microarray-based gene expression analysis holds promise of improving

prognostication and treatment decisions for breast cancer patients. However, the

heterogeneity of breast cancer emphasizes the need for validation of prognostic

gene signatures in larger sample sets stratified into relevant subgroups. Here,

we describe a multifunctional user-friendly online tool, GOBO

(http://co.bmc.lu.se/gobo), allowing a range of different analyses to be

performed in an 1881-sample breast tumor data set, and a 51-sample breast cancer

cell line set, both generated on Affymetrix U133A microarrays. GOBO supports a

wide range of applications including: 1) rapid assessment of gene expression

levels in subgroups of breast tumors and cell lines, 2) identification of

co-expressed genes for creation of potential metagenes, 3) association with

outcome for gene expression levels of single genes, sets of genes, or gene

signatures in multiple subgroups of the 1881-sample breast cancer data set. The

design and implementation of GOBO facilitate easy incorporation of additional

query functions and applications, as well as additional data sets irrespective of

tumor type and array platform.

DOI: 10.1371/journal.pone.0017911

PMCID: PMC3061871

PMID: 21445301 [Indexed for MEDLINE]

1744. BMC Bioinformatics. 2011 Mar 17;12:76. doi: 10.1186/1471-2105-12-76.

Outer membrane proteins can be simply identified using secondary structure

element alignment.

Yan RX(1), Chen Z, Zhang Z.

Author information:

(1)State Key Laboratory of Agrobiotechnology, College of Biological Sciences,

China Agricultural University, Beijing, PR China.

BACKGROUND: Outer membrane proteins (OMPs) are frequently found in the outer

membranes of gram-negative bacteria, mitochondria and chloroplasts and have been

found to play diverse functional roles. Computational discrimination of OMPs from

globular proteins and other types of membrane proteins is helpful to accelerate

new genome annotation and drug discovery.

RESULTS: Based on the observation that almost all OMPs consist of antiparallel

β-strands in a barrel shape and that their secondary structure arrangements

differ from those of other types of proteins, we propose a simple method called

SSEA-OMP to identify OMPs using secondary structure element alignment. Through

intensive benchmark experiments, the proposed SSEA-OMP method is better than some

well-established OMP detection methods.

CONCLUSIONS: The major advantage of SSEA-OMP is its good prediction performance

considering its simplicity. The web server implements the method is freely

accessible at http://protein.cau.edu.cn/SSEA-OMP/index.html.

DOI: 10.1186/1471-2105-12-76

PMCID: PMC3072342

PMID: 21414186 [Indexed for MEDLINE]

1745. Bioinformatics. 2011 Mar 15;27(6):889-90. doi: 10.1093/bioinformatics/btr020.

Epub 2011 Jan 19.

Dalliance: interactive genome viewing on the web.

Down TA(1), Piipari M, Hubbard TJ.

Author information:

(1)Wellcome Trust/CRUK Gurdon Institute, Cambridge CB2 1QN, UK.

thomas@biodalliance.org

SUMMARY: Dalliance is a new genome viewer which offers a high level of

interactivity while running within a web browser. All data is fetched using the

established distributed annotation system (DAS) protocol, making it easy to

customize the browser and add extra data.

AVAILABILITY AND IMPLEMENTATION: Dalliance runs entirely within your web browser,

and relies on existing DAS server infrastructure. Browsers for several mammalian

genomes are available at http://www.biodalliance.org/, and the use of DAS means

you can add your own data to these browsers. In addition, the source code

(Javascript) is available under the BSD license, and is straightforward to

install on your own web server and embed within other documents.

DOI: 10.1093/bioinformatics/btr020

PMCID: PMC3051325

PMID: 21252075 [Indexed for MEDLINE]

1746. Bioinformatics. 2011 Mar 15;27(6):771-6. doi: 10.1093/bioinformatics/btr016. Epub

2011 Jan 11.

A k-mer scheme to predict piRNAs and characterize locust piRNAs.

Zhang Y(1), Wang X, Kang L.

Author information:

(1)Institute of Zoology, Chinese Academy of Sciences, Beijing, China.

MOTIVATION: Identifying piwi-interacting RNAs (piRNAs) of non-model organisms is

a difficult and unsolved problem because piRNAs lack conservative secondary

structure motifs and sequence homology in different species.

RESULTS: In this article, a k-mer scheme is proposed to identify piRNA sequences,

relying on the training sets from non-piRNA and piRNA sequences of five model

species sequenced: rat, mouse, human, fruit fly and nematode. Compared with the

existing 'static' scheme based on the position-specific base usage, our novel

'dynamic' algorithm performs much better with a precision of over 90% and a

sensitivity of over 60%, and the precision is verified by 5-fold cross-validation

in these species. To test its validity, we use the algorithm to identify piRNAs

of the migratory locust based on 603 607 deep-sequenced small RNA sequences.

Totally, 87 536 piRNAs of the locust are predicted, and 4426 of them matched with

existing locust transposons. The transcriptional difference between solitary and

gregarious locusts was described. We also revisit the position-specific base

usage of piRNAs and find the conservation in the end of piRNAs. Therefore, the

method we developed can be used to identify piRNAs of non-model organisms without

complete genome sequences.

AVAILABILITY: The web server for implementing the algorithm and the software code

are freely available to the academic community at

http://59.79.168.90/piRNA/index.php.

DOI: 10.1093/bioinformatics/btr016

PMCID: PMC3051322

PMID: 21224287 [Indexed for MEDLINE]

1747. BMC Bioinformatics. 2011 Mar 9;12:71. doi: 10.1186/1471-2105-12-71.

LabKey Server: an open source platform for scientific data integration, analysis

and collaboration.

Nelson EK(1), Piehler B, Eckels J, Rauch A, Bellew M, Hussey P, Ramsay S, Nathe

C, Lum K, Krouse K, Stearns D, Connolly B, Skillman T, Igra M.

Author information:

(1)LabKey Software, Seattle, WA 98102, USA. eknelson@labkey.com

BACKGROUND: Broad-based collaborations are becoming increasingly common among

disease researchers. For example, the Global HIV Enterprise has united

cross-disciplinary consortia to speed progress towards HIV vaccines through

coordinated research across the boundaries of institutions, continents and

specialties. New, end-to-end software tools for data and specimen management are

necessary to achieve the ambitious goals of such alliances. These tools must

enable researchers to organize and integrate heterogeneous data early in the

discovery process, standardize processes, gain new insights into pooled data and

collaborate securely.

RESULTS: To meet these needs, we enhanced the LabKey Server platform, formerly

known as CPAS. This freely available, open source software is maintained by

professional engineers who use commercially proven practices for software

development and maintenance. Recent enhancements support: (i) Submitting

specimens requests across collaborating organizations (ii) Graphically defining

new experimental data types, metadata and wizards for data collection (iii)

Transitioning experimental results from a multiplicity of spreadsheets to custom

tables in a shared database (iv) Securely organizing, integrating, analyzing,

visualizing and sharing diverse data types, from clinical records to specimens to

complex assays (v) Interacting dynamically with external data sources (vi)

Tracking study participants and cohorts over time (vii) Developing custom

interfaces using client libraries (viii) Authoring custom visualizations in a

built-in R scripting environment. Diverse research organizations have adopted and

adapted LabKey Server, including consortia within the Global HIV Enterprise.

Atlas is an installation of LabKey Server that has been tailored to serve these

consortia. It is in production use and demonstrates the core capabilities of

LabKey Server. Atlas now has over 2,800 active user accounts originating from

approximately 36 countries and 350 organizations. It tracks roughly 27,000 assay

runs, 860,000 specimen vials and 1,300,000 vial transfers.

CONCLUSIONS: Sharing data, analysis tools and infrastructure can speed the

efforts of large research consortia by enhancing efficiency and enabling new

insights. The Atlas installation of LabKey Server demonstrates the utility of the

LabKey platform for collaborative research. Stable, supported builds of LabKey

Server are freely available for download at http://www.labkey.org. Documentation

and source code are available under the Apache License 2.0.

DOI: 10.1186/1471-2105-12-71

PMCID: PMC3062597

PMID: 21385461 [Indexed for MEDLINE]

1748. BMC Res Notes. 2011 Mar 9;4:57. doi: 10.1186/1756-0500-4-57.

QiSampler: evaluation of scoring schemes for high-throughput datasets using a

repetitive sampling strategy on gold standards.

Fontaine JF(1), Suter B, Andrade-Navarro MA.

Author information:

(1)Max Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125

Berlin, Germany. jean-fred.fontaine@mdc-berlin.de.

BACKGROUND: High-throughput biological experiments can produce a large amount of

data showing little overlap with current knowledge. This may be a problem when

evaluating alternative scoring mechanisms for such data according to a gold

standard dataset because standard statistical tests may not be appropriate.

FINDINGS: To address this problem we have implemented the QiSampler tool that

uses a repetitive sampling strategy to evaluate several scoring schemes or

experimental parameters for any type of high-throughput data given a gold

standard. We provide two example applications of the tool: selection of the best

scoring scheme for a high-throughput protein-protein interaction dataset by

comparison to a dataset derived from the literature, and evaluation of functional

enrichment in a set of tumour-related differentially expressed genes from a

thyroid microarray dataset.

CONCLUSIONS: QiSampler is implemented as an open source R script and a web

server, which can be accessed at http://cbdm.mdc-berlin.de/tools/sampler/.

DOI: 10.1186/1756-0500-4-57

PMCID: PMC3060832

PMID: 21388526

1749. Amino Acids. 2011 Mar;40(3):963-73. doi: 10.1007/s00726-010-0721-1. Epub 2010 Aug

21.

iFC²: an integrated web-server for improved prediction of protein structural

class, fold type, and secondary structure content.

Chen K(1), Stach W, Homaeian L, Kurgan L.

Author information:

(1)Department of Electrical and Computer Engineering, University of Alberta,

Edmonton, Alberta, Canada.

Several descriptors of protein structure at the sequence and residue levels have

been recently proposed. They are widely adopted in the analysis and prediction of

structural and functional characteristics of proteins. Numerous in silico methods

have been developed for sequence-based prediction of these descriptors. However,

many of them do not have a public web-server and only a few integrate multiple

descriptors to improve the predictions. We introduce iFC² (integrated prediction

of fold, class, and content) server that is the first to integrate three modern

predictors of sequence-level descriptors. They concern fold type (PFRES),

structural class (SCEC), and secondary structure content (PSSC-core). The server

exploits relations between the three descriptors to implement a cross-evaluation

procedure that improves over the predictions of the individual methods. The iFC²

annotates fold and class predictions as potentially correct/incorrect. When

tested on datasets with low-similarity chains, for the fold prediction iFC²

labels 82% of the PFRES predictions as correct and the accuracy of these

predictions equals 72%. The accuracy of the remaining 28% of the PFRES

predictions equals 38%. Similarly, our server assigns correct labels for over 79%

of SCEC predictions, which are shown to be 98% accurate, while the remaining SCEC

predictions are only 15% accurate. These results are shown to be competitive when

contrasted against recent relevant web-servers. Predictions on CASP8 targets show

that the content predicted by iFC² is competitive when compared with the content

computed from the tertiary structures predicted by three best-performing methods

in CASP8. The iFC² server is available at

http://biomine.ece.ualberta.ca/1D/1D.html .

DOI: 10.1007/s00726-010-0721-1

PMID: 20730460 [Indexed for MEDLINE]

1750. Bioinformatics. 2011 Mar 1;27(5):720-2. doi: 10.1093/bioinformatics/btq722. Epub

2011 Jan 17.

Identifying viral integration sites using SeqMap 2.0.

Hawkins TB(1), Dantzer J, Peters B, Dinauer M, Mockaitis K, Mooney S, Cornetta K.

Author information:

(1)Department of Medical and Molecular Genetics, Indiana University School of

Medicine, Indianapolis, IN 46202, USA. troyhawk@iupui.edu

Retroviral integration has been implicated in several biomedical applications,

including identification of cancer-associated genes and malignant transformation

in gene therapy clinical trials. We introduce an efficient and scalable method

for fast identification of viral vector integration sites from long read

high-throughput sequencing. Individual sequence reads are masked to remove

non-genomic sequence, aligned to the host genome and assembled into contiguous

fragments used to pinpoint the position of integration.AVAILABILITY AND

IMPLEMENTATION: The method is implemented in a publicly accessible web server

platform, SeqMap 2.0, containing analysis tools and both private and shared lab

workspaces that facilitate collaboration among researchers. Available at

http://seqmap.compbio.iupui.edu/.

DOI: 10.1093/bioinformatics/btq722

PMCID: PMC3042184

PMID: 21245052 [Indexed for MEDLINE]

1751. Bioinformatics. 2011 Mar 1;27(5):730-1. doi: 10.1093/bioinformatics/btr001. Epub

2011 Jan 5.

MBRole: enrichment analysis of metabolomic data.

Chagoyen M(1), Pazos F.

Author information:

(1)Computational Systems Biology Group, National Centre for Biotechnology

(CNB-CSIC), Madrid, Spain. monica.chagoyen@cnb.csic.es

While many tools exist for performing enrichment analysis of transcriptomic and

proteomic data in order to interpret them in biological terms, almost no

equivalent tools exist for metabolomic data. We present Metabolite Biological

Role (MBRole), a web server for carrying out over-representation analysis of

biological and chemical annotations in arbitrary sets of metabolites (small

chemical compounds) coming from metabolomic data of any organism or

sample.AVAILABILITY AND IMPLEMENTATION: The web server is freely available at

http://csbg.cnb.csic.es/mbrole. It was tested in the main web browsers.

DOI: 10.1093/bioinformatics/btr001

PMID: 21208985 [Indexed for MEDLINE]

1752. J Mol Recognit. 2011 Mar-Apr;24(2):303-13. doi: 10.1002/jmr.1061.

SVM based prediction of RNA-binding proteins using binding residues and

evolutionary information.

Kumar M(1), Gromiha MM, Raghava GP.

Author information:

(1)Department of Biology, McGill University, Montreal, QC, H3A 1B1, Canada.

RNA-binding proteins (RBPs) play crucial role in transcription and

gene-regulation. This paper describes a support vector machine (SVM) based method

for discriminating and classifying RNA-binding and non-binding proteins using

sequence features. With the threshold of 30% interacting residues, RNA-binding

amino acid prediction method PPRINT achieved the Matthews correlation coefficient

(MCC) of 0.32. BLAST and PSI-BLAST identified RBPs with the coverage of 32.63 and

33.16%, respectively, at the e-value of 1e-4. The SVM models developed with amino

acid, dipeptide and four-part amino acid compositions showed the MCC of 0.60,

0.46, and 0.53, respectively. This is the first study in which evolutionary

information in form of position specific scoring matrix (PSSM) profile has been

successfully used for predicting RBPs. We achieved the maximum MCC of 0.62 using

SVM model based on PSSM called PSSM-400. Finally, we developed different hybrid

approaches and achieved maximum MCC of 0.66. We also developed a method for

predicting three subclasses of RNA binding proteins (e.g., rRNA, tRNA, mRNA

binding proteins). The performance of the method was also evaluated on an

independent dataset of 69 RBPs and 100 non-RBPs (NBPs). An additional

benchmarking was also performed using gene ontology (GO) based annotation. Based

on the hybrid approach a web-server RNApred has been developed for predicting RNA

binding proteins from amino acid sequences

(http://www.imtech.res.in/raghava/rnapred/).

Copyright © 2010 John Wiley & Sons, Ltd.

DOI: 10.1002/jmr.1061

PMID: 20677174 [Indexed for MEDLINE]

1753. J Struct Biol. 2011 Mar;173(3):558-69. doi: 10.1016/j.jsb.2010.09.009. Epub 2010

Sep 17.

The utility of geometrical and chemical restraint information extracted from

predicted ligand-binding sites in protein structure refinement.

Brylinski M(1), Lee SY, Zhou H, Skolnick J.

Author information:

(1)Center for the Study of Systems Biology, Georgia Institute of Technology,

Atlanta, GA 30318, USA.

Exhaustive exploration of molecular interactions at the level of complete

proteomes requires efficient and reliable computational approaches to protein

function inference. Ligand docking and ranking techniques show considerable

promise in their ability to quantify the interactions between proteins and small

molecules. Despite the advances in the development of docking approaches and

scoring functions, the genome-wide application of many ligand docking/screening

algorithms is limited by the quality of the binding sites in theoretical receptor

models constructed by protein structure prediction. In this study, we describe a

new template-based method for the local refinement of ligand-binding regions in

protein models using remotely related templates identified by threading. We

designed a Support Vector Regression (SVR) model that selects correct binding

site geometries in a large ensemble of multiple receptor conformations. The SVR

model employs several scoring functions that impose geometrical restraints on the

Cα positions, account for the specific chemical environment within a binding site

and optimize the interactions with putative ligands. The SVR score is well

correlated with the RMSD from the native structure; in 47% (70%) of the cases,

the Pearson's correlation coefficient is >0.5 (>0.3). When applied to weakly

homologous models, the average heavy atom, local RMSD from the native structure

of the top-ranked (best of top five) binding site geometries is 3.1Å (2.9Å) for

roughly half of the targets; this represents a 0.1 (0.3)Å average improvement

over the original predicted structure. Focusing on the subset of strongly

conserved residues, the average heavy atom RMSD is 2.6Å (2.3Å). Furthermore, we

estimate the upper bound of template-based binding site refinement using only

weakly related proteins to be ∼2.6Å RMSD. This value also corresponds to the

plasticity of the ligand-binding regions in distant homologues. The Binding Site

Refinement (BSR) approach is available to the scientific community as a web

server that can be accessed at http://cssb.biology.gatech.edu/bsr/.

Copyright © 2010 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.jsb.2010.09.009

PMCID: PMC3036769

PMID: 20850544 [Indexed for MEDLINE]

1754. J Struct Funct Genomics. 2011 Mar;12(1):33-41. doi: 10.1007/s10969-011-9108-0.

Epub 2011 Apr 26.

PRICE (PRotein Interface Conservation and Energetics): a server for the analysis

of protein-protein interfaces.

Guharoy M(1), Pal A, Dasgupta M, Chakrabarti P.

Author information:

(1)Department of Biochemistry, Bose Institute, Kolkata, India.

Residues in a protein-protein interface that are important for forming and

stabilizing the interaction can usually be identified by looking at patterns of

evolutionary conservation in groups of homologous proteins and also by the

computational identification of binding hotspots. The PRICE (PRotein Interface

Conservation and Energetics) server takes the coordinates of a protein-protein

complex, dissects the interface into core and rim regions, and calculates (1) the

degree of conservation (measured as the sequence entropy), as well as (2) the

change in free energy of binding (∆∆G, due to alanine scanning mutagenesis) of

interface residues. Results are displayed as color-coded plots and also made

available for download. This enables the computational identification of binding

hot spots, based on which further experiments can be designed. The method will

aid in protein functional prediction by correct assignment of hot regions

involved in binding. Consideration of sequence entropies for residues with large

∆∆G values may provide an indication of the biological relevance of the

interface. Finally, the results obtained on a test set of alanine mutants has

been compared to those obtained using other servers/methods. The PRICE server is

a web application available at

http://www.boseinst.ernet.in/resources/bioinfo/stag.html.

DOI: 10.1007/s10969-011-9108-0

PMID: 21519818 [Indexed for MEDLINE]

1755. Mitochondrion. 2011 Mar;11(2):351-6. doi: 10.1016/j.mito.2010.09.013. Epub 2010

Oct 7.

MitoTool: a web server for the analysis and retrieval of human mitochondrial DNA

sequence variations.

Fan L(1), Yao YG.

Author information:

(1)Key Laboratory of Animal Models and Human Disease Mechanisms of the Chinese

Academy of Sciences & Yunnan Province, Kunming Institute of Zoology, Kunming,

Yunnan, China. mitotool@gmail.com

MitoTool, a web-based bioinformatics platform, is designed for deciphering human

mitochondrial DNA (mtDNA) data in batch mode. The platform has advantages in (i)

parsing diverse types of mtDNA data; (ii) automatically classifying haplogroup

according to mtDNA sequences or variants; (iii) discovering possibly missing

variants of the samples with claimed haplogroups status; (iv) estimating the

evolutionary conservation index, protein coding effect and potential

pathogenicity of certain substitutions; (v) performing statistical analysis for

haplogroup distribution frequency between case and control groups. Furthermore,

it offers an integrated database for retrieving five types of

mitochondrion-related information. The MitoTool is freely accessed at

http://www.mitotool.org.

Copyright © 2010. Published by Elsevier B.V.

DOI: 10.1016/j.mito.2010.09.013

PMID: 20933105 [Indexed for MEDLINE]

1756. Mol Biosyst. 2011 Mar;7(3):911-9. doi: 10.1039/c0mb00170h. Epub 2010 Dec 23.

GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing

two different modes of pseudo amino acid compositions.

Xiao X(1), Wang P, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403,

China. xxiao@gordonlifescience.org

G protein-coupled receptors (GPCRs) are among the most frequent targets of

therapeutic drugs. With the avalanche of newly generated protein sequences in the

post genomic age, to expedite the process of drug discovery, it is highly

desirable to develop an automated method to rapidly identify GPCRs and their

types. A new predictor was developed by hybridizing two different modes of

pseudo-amino acid composition (PseAAC): the functional domain PseAAC and the

low-frequency Fourier spectrum PseAAC. The new predictor is called GPCR-2L, where

"2L" means that it is a two-layer predictor: the 1st layer prediction engine is

to identify a query protein as GPCR or not; if it is, the prediction will be

automatically continued to further identify it as belonging to one of the

following six types: (1) rhodopsin-like (Class A), (2) secretin-like (Class B),

(3) metabotropic glutamate/pheromone (Class C), (4) fungal pheromone (Class D),

(5) cAMP receptor (Class E), or (6) frizzled/smoothened family (Class F). The

overall success rate of GPCR-2L in identifying proteins as GPCRs or non-GPCRs is

over 97.2%, while identifying GPCRs among their six types is over 97.8%. Such

high success rates were derived by the rigorous jackknife cross-validation on a

stringent benchmark dataset, in which none of the included proteins had ≥40%

pairwise sequence identity to any other protein in a same subset. As a

user-friendly web-server, GPCR-2L is freely accessible to the public at

http://icpr.jci.edu.cn/, by which one can obtain the 2-level results in about 20

s for a query protein sequence of 500 amino acids. The longer the sequence is,

the more time it may usually need. The high success rates reported here indicate

that it is a quite effective approach to identify GPCRs and their types with the

functional domain information and the low-frequency Fourier spectrum analysis. It

is anticipated that GPCR-2L may become a useful tool for both basic research and

drug development in the areas related to GPCRs.

DOI: 10.1039/c0mb00170h

PMID: 21180772 [Indexed for MEDLINE]

1757. Nucleic Acids Res. 2011 Mar;39(4):1187-96. doi: 10.1093/nar/gkq958. Epub 2010 Oct

20.

Identification of new homologs of PD-(D/E)XK nucleases by support vector machines

trained on data derived from profile-profile alignments.

Laganeckas M(1), Margelevicius M, Venclovas C.

Author information:

(1)Institute of Biotechnology, Graičiūno 8, LT-02241 Vilnius, Lithuania.

PD-(D/E)XK nucleases, initially represented by only Type II restriction enzymes,

now comprise a large and extremely diverse superfamily of proteins. They

participate in many different nucleic acids transactions including DNA

degradation, recombination, repair and RNA processing. Different PD-(D/E)XK

families, although sharing a structurally conserved core, typically display

little or no detectable sequence similarity except for the active site motifs.

This makes the identification of new superfamily members using standard homology

search techniques challenging. To tackle this problem, we developed a method for

the detection of PD-(D/E)XK families based on the binary classification of

profile-profile alignments using support vector machines (SVMs). Using a number

of both superfamily-specific and general features, SVMs were trained to identify

true positive alignments of PD-(D/E)XK representatives. With this method we

identified several PFAM families of uncharacterized proteins as putative new

members of the PD-(D/E)XK superfamily. In addition, we assigned several

unclassified restriction enzymes to the PD-(D/E)XK type. Results show that the

new method is able to make confident assignments even for alignments that have

statistically insignificant scores. We also implemented the method as a freely

accessible web server at http://www.ibt.lt/bioinformatics/software/pdexk/.

DOI: 10.1093/nar/gkq958

PMCID: PMC3045609

PMID: 20961958 [Indexed for MEDLINE]

1758. PLoS One. 2011 Feb 28;6(2):e16774. doi: 10.1371/journal.pone.0016774.

Predictions of hot spot residues at protein-protein interfaces using support

vector machines.

Lise S(1), Buchan D, Pontil M, Jones DT.

Author information:

(1)Department of Computer Science, University College London, London, United

Kingdom.

Protein-protein interactions are critically dependent on just a few 'hot spot'

residues at the interface. Hot spots make a dominant contribution to the free

energy of binding and they can disrupt the interaction if mutated to alanine.

Here, we present HSPred, a support vector machine(SVM)-based method to predict

hot spot residues, given the structure of a complex. HSPred represents an

improvement over a previously described approach (Lise et al, BMC Bioinformatics

2009, 10:365). It achieves higher accuracy by treating separately predictions

involving either an arginine or a glutamic acid residue. These are the amino acid

types on which the original model did not perform well. We have therefore

developed two additional SVM classifiers, specifically optimised for these cases.

HSPred reaches an overall precision and recall respectively of 61% and 69%, which

roughly corresponds to a 10% improvement. An implementation of the described

method is available as a web server at http://bioinf.cs.ucl.ac.uk/hspred. It is

free to non-commercial users.

DOI: 10.1371/journal.pone.0016774

PMCID: PMC3046169

PMID: 21386962 [Indexed for MEDLINE]

1759. Bioinformatics. 2011 Feb 1;27(3):426-7. doi: 10.1093/bioinformatics/btq664. Epub

2010 Dec 11.

Improved predictions by Pcons.net using multiple templates.

Larsson P(1), Skwark MJ, Wallner B, Elofsson A.

Author information:

(1)Center for Biomembrane Research, Department of Biochemistry and Biophysics,

Swedish e-Science Research Centre, Stockholm Bioinformatics Centre, SciLifeLab,

Stockholm University SE-10691 Stockholm, Sweden.

Multiple templates can often be used to build more accurate homology models than

models built from a single template. Here we introduce PconsM, an automated

protocol that uses multiple templates to build protein models. PconsM has been

among the top-performing methods in the recent CASP experiments and consistently

perform better than the single template models used in Pcons.net. In particular

for the easier targets with many alternative templates with a high degree of

sequence identity, quality is readily improved with a few percentages over the

highest ranked model built on a single template. PconsM is available as an

additional pipeline within the Pcons.net protein structure prediction

server.AVAILABILITY AND IMPLEMENTATION: PconsM is freely available from

http://pcons.net/.

DOI: 10.1093/bioinformatics/btq664

PMCID: PMC3031036

PMID: 21149277 [Indexed for MEDLINE]

1760. Bioinformatics. 2011 Feb 1;27(3):343-50. doi: 10.1093/bioinformatics/btq662. Epub

2010 Dec 5.

Toward the estimation of the absolute quality of individual protein structure

models.

Benkert P(1), Biasini M, Schwede T.

Author information:

(1)Biozentrum, University of Basel, Switzerland.

MOTIVATION: Quality assessment of protein structures is an important part of

experimental structure validation and plays a crucial role in protein structure

prediction, where the predicted models may contain substantial errors. Most

current scoring functions are primarily designed to rank alternative models of

the same sequence supporting model selection, whereas the prediction of the

absolute quality of an individual protein model has received little attention in

the field. However, reliable absolute quality estimates are crucial to assess the

suitability of a model for specific biomedical applications.

RESULTS: In this work, we present a new absolute measure for the quality of

protein models, which provides an estimate of the 'degree of nativeness' of the

structural features observed in a model and describes the likelihood that a given

model is of comparable quality to experimental structures. Model quality

estimates based on the QMEAN scoring function were normalized with respect to the

number of interactions. The resulting scoring function is independent of the size

of the protein and may therefore be used to assess both monomers and entire

oligomeric assemblies. Model quality scores for individual models are then

expressed as 'Z-scores' in comparison to scores obtained for high-resolution

crystal structures. We demonstrate the ability of the newly introduced QMEAN

Z-score to detect experimentally solved protein structures containing significant

errors, as well as to evaluate theoretical protein models. In a comprehensive

QMEAN Z-score analysis of all experimental structures in the PDB, membrane

proteins accumulate on one side of the score spectrum and thermostable proteins

on the other. Proteins from the thermophilic organism Thermatoga maritima

received significantly higher QMEAN Z-scores in a pairwise comparison with their

homologous mesophilic counterparts, underlining the significance of the QMEAN

Z-score as an estimate of protein stability.

AVAILABILITY: The Z-score calculation has been integrated in the QMEAN server

available at: http://swissmodel.expasy.org/qmean.

DOI: 10.1093/bioinformatics/btq662

PMCID: PMC3031035

PMID: 21134891 [Indexed for MEDLINE]

1761. Bioinformatics. 2011 Feb 1;27(3):317-25. doi: 10.1093/bioinformatics/btq651. Epub

2010 Dec 1.

maxAlike: maximum likelihood-based sequence reconstruction with application to

improved primer design for unknown sequences.

Menzel P(1), Stadler PF, Gorodkin J.

Author information:

(1)Center for non-coding RNA in Technology and Health, IBHV, University of

Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark.

MOTIVATION: The task of reconstructing a genomic sequence from a particular

species is gaining more and more importance in the light of the rapid development

of high-throughput sequencing technologies and their limitations. Applications

include not only compensation for missing data in unsequenced genomic regions and

the design of oligonucleotide primers for target genes in species with lacking

sequence information but also the preparation of customized queries for homology

searches.

RESULTS: We introduce the maxAlike algorithm, which reconstructs a genomic

sequence for a specific taxon based on sequence homologs in other species. The

input is a multiple sequence alignment and a phylogenetic tree that also contains

the target species. For this target species, the algorithm computes nucleotide

probabilities at each sequence position. Consensus sequences are then

reconstructed based on a certain confidence level. For 37 out of 44 target

species in a test dataset, we obtain a significant increase of the reconstruction

accuracy compared to both the consensus sequence from the alignment and the

sequence of the nearest phylogenetic neighbor. When considering only nucleotides

above a confidence limit, maxAlike is significantly better (up to 10%) in all 44

species. The improved sequence reconstruction also leads to an increase of the

quality of PCR primer design for yet unsequenced genes: the differences between

the expected T(m) and real T(m) of the primer-template duplex can be reduced by

~26% compared with other reconstruction approaches. We also show that the

prediction accuracy is robust to common distortions of the input trees. The

prediction accuracy drops by only 1% on average across all species for 77% of

trees derived from random genomic loci in a test dataset.

AVAILABILITY: maxAlike is available for download and web server at:

http://rth.dk/resources/maxAlike.

DOI: 10.1093/bioinformatics/btq651

PMCID: PMC3031029

PMID: 21123221 [Indexed for MEDLINE]

1762. J Bioinform Comput Biol. 2011 Feb;9(1):43-65.

Prediction of the exposure status of transmembrane beta barrel residues from

protein sequence.

Hayat S(1), Walter P, Park Y, Helms V.

Author information:

(1)Center for Bioinformatics, Saarland University, Germany.

s.hayat@bioinformatik.uni-saarland.de

We present BTMX (Beta barrel TransMembrane eXposure), a computational method to

predict the exposure status (i.e. exposed to the bilayer or hidden in the protein

structure) of transmembrane residues in transmembrane beta barrel proteins

(TMBs). BTMX predicts the exposure status of known TM residues with an accuracy

of 84.2% over 2,225 residues and provides a confidence score for all predictions.

Predictions made are in concert with the fact that hydrophobic residues tend to

be more exposed to the bilayer. The biological relevance of the input parameters

is also discussed. The highest prediction accuracy is obtained when a sliding

window comprising three residues with similar C(α)-C(β) vector orientations is

employed. The prediction accuracy of the BTMX method on a separate unseen

non-redundant test dataset is 78.1%. By employing out-pointing residues that are

exposed to the bilayer, we have identified various physico-chemical properties

that show statistically significant differences between the beta strands located

at the oligomeric interfaces compared to the non-oligomeric strands. The BTMX web

server generates colored, annotated snake-plots as part of the prediction results

and is available under the BTMX tab at

http://service.bioinformatik.uni-saarland.de/tmx-site/. Exposure status

prediction of TMB residues may be useful in 3D structure prediction of TMBs.

PMID: 21328706 [Indexed for MEDLINE]

1763. J Bioinform Comput Biol. 2011 Feb;9(1):15-41.

Error tolerant NMR backbone resonance assignment and automated structure

generation.

Alipanahi B(1), Gao X, Karakoc E, Li SC, Balbach F, Feng G, Donaldson L, Li M.

Author information:

(1)David R. Cheriton School of Computer Science, University of Waterloo,

Waterloo, Ontario N2L3G1, Canada.

Error tolerant backbone resonance assignment is the cornerstone of the NMR

structure determination process. Although a variety of assignment approaches have

been developed, none works sufficiently well on noisy fully automatically picked

peaks to enable the subsequent automatic structure determination steps. We have

designed an integer linear programming (ILP) based assignment system (IPASS) that

has enabled fully automatic protein structure determination for four test

proteins. IPASS employs probabilistic spin system typing based on chemical shifts

and secondary structure predictions. Furthermore, IPASS extracts connectivity

information from the inter-residue information and the (automatically picked)

(15)N-edited NOESY peaks which are then used to fix reliable fragments. When

applied to automatically picked peaks for real proteins, IPASS achieves an

average precision and recall of 82% and 63%, respectively. In contrast, the next

best method, MARS, achieves an average precision and recall of 77% and 36%,

respectively. The assignments generated by IPASS are then fed into our protein

structure calculation system, FALCON-NMR, to determine the 3D structures without

human intervention. The final models have backbone RMSDs of 1.25Å, 0.88Å, 1.49Å,

and 0.67Å to the reference native structures for proteins TM1112, CASKIN, VRAR,

and HACS1, respectively. The web server is publicly available at

http://monod.uwaterloo.ca/nmr/ipass.

PMID: 21328705 [Indexed for MEDLINE]

1764. Mol Divers. 2011 Feb;15(1):149-55. doi: 10.1007/s11030-010-9227-8. Epub 2010 Feb

11.

Quat-2L: a web-server for predicting protein quaternary structural attributes.

Xiao X(1), Wang P, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333001,

China. xiaoxuan0326@yahoo.com.cn

By hybridizing the functional-domain and sequence-correlated pseudo amino acid

composition approaches, a 2-layer predictor called "Quat-2L" was developed for

predicting the quaternary structural attribute of a protein according to its

sequence information alone. The 1st layer is to identify the query protein as

monomer, homo-oligomer, or hetero-oligomer. If the result thus obtained turns out

to be homo-oligomer or hetero-oligomer, then the prediction will be automatically

continued to further identify it belonging to one of the following six subtypes:

(1) dimer, (2) trimer, (3) tetramer, (4) pentamer, (5) hexamer, and (6) octamer.

The overall success rate of Quat-2L for the 1st layer identification was 71.14%;

while the overall success rates of the 2nd layer for homo-oligomers and

hetero-oligomers were 76.91 and 82.52%, respectively. These rates were derived by

the jackknife cross-validation tests on the stringent benchmark data set in which

none of proteins has ≥ 60% pairwise sequence identity to any other in the same

subset. As a web-server, Quat-2L is freely accessible to the public via

http://icpr.jci.jx.cn/bioinfo/Quat-2L, where one can get 2-level results in about

15 s.

DOI: 10.1007/s11030-010-9227-8

PMID: 20148364 [Indexed for MEDLINE]

1765. PLoS One. 2011 Jan 28;6(1):e16178. doi: 10.1371/journal.pone.0016178.

Computing the partition function for kinetically trapped RNA secondary

structures.

Lorenz WA(1), Clote P.

Author information:

(1)Department of Mathematics and Computer Science, Denison University, Granville,

Ohio, United States of America.

An RNA secondary structure is locally optimal if there is no lower energy

structure that can be obtained by the addition or removal of a single base pair,

where energy is defined according to the widely accepted Turner nearest neighbor

model. Locally optimal structures form kinetic traps, since any evolution away

from a locally optimal structure must involve energetically unfavorable folding

steps. Here, we present a novel, efficient algorithm to compute the partition

function over all locally optimal secondary structures of a given RNA sequence.

Our software, RNAlocopt runs in O(n3) time and O(n2) space. Additionally,

RNAlocopt samples a user-specified number of structures from the Boltzmann

subensemble of all locally optimal structures. We apply RNAlocopt to show that

(1) the number of locally optimal structures is far fewer than the total number

of structures--indeed, the number of locally optimal structures approximately

equal to the square root of the number of all structures, (2) the structural

diversity of this subensemble may be either similar to or quite different from

the structural diversity of the entire Boltzmann ensemble, a situation that

depends on the type of input RNA, (3) the (modified) maximum expected accuracy

structure, computed by taking into account base pairing frequencies of locally

optimal structures, is a more accurate prediction of the native structure than

other current thermodynamics-based methods. The software RNAlocopt constitutes a

technical breakthrough in our study of the folding landscape for RNA secondary

structures. For the first time, locally optimal structures (kinetic traps in the

Turner energy model) can be rapidly generated for long RNA sequences, previously

impossible with methods that involved exhaustive enumeration. Use of locally

optimal structure leads to state-of-the-art secondary structure prediction, as

benchmarked against methods involving the computation of minimum free energy and

of maximum expected accuracy. Web server and source code available at

http://bioinformatics.bc.edu/clotelab/RNAlocopt/.

DOI: 10.1371/journal.pone.0016178

PMCID: PMC3030561

PMID: 21297972 [Indexed for MEDLINE]

1766. Algorithms Mol Biol. 2011 Jan 24;6:2. doi: 10.1186/1748-7188-6-2.

WordCluster: detecting clusters of DNA words and genomic elements.

Hackenberg M(1), Carpena P, Bernaola-Galván P, Barturen G, Alganza AM, Oliver JL.

Author information:

(1)Dpto, de Genética, Facultad de Ciencias, Universidad de Granada, Campus de

Fuentenueva s/n, 18071-Granada & Lab, de Bioinformática, Centro de Investigación

Biomédica, PTS, Avda, del Conocimiento s/n, 18100-Granada, Spain.

mlhack@gmail.com.

BACKGROUND: Many k-mers (or DNA words) and genomic elements are known to be

spatially clustered in the genome. Well established examples are the genes,

TFBSs, CpG dinucleotides, microRNA genes and ultra-conserved non-coding regions.

Currently, no algorithm exists to find these clusters in a statistically

comprehensible way. The detection of clustering often relies on densities and

sliding-window approaches or arbitrarily chosen distance thresholds.

RESULTS: We introduce here an algorithm to detect clusters of DNA words (k-mers),

or any other genomic element, based on the distance between consecutive copies

and an assigned statistical significance. We implemented the method into a web

server connected to a MySQL backend, which also determines the co-localization

with gene annotations. We demonstrate the usefulness of this approach by

detecting the clusters of CAG/CTG (cytosine contexts that can be methylated in

undifferentiated cells), showing that the degree of methylation vary drastically

between inside and outside of the clusters. As another example, we used

WordCluster to search for statistically significant clusters of olfactory

receptor (OR) genes in the human genome.

CONCLUSIONS: WordCluster seems to predict biological meaningful clusters of DNA

words (k-mers) and genomic entities. The implementation of the method into a web

server is available at http://bioinfo2.ugr.es/wordCluster/wordCluster.php

including additional features like the detection of co-localization with gene

regions or the annotation enrichment tool for functional analysis of overlapped

genes.

DOI: 10.1186/1748-7188-6-2

PMCID: PMC3037320

PMID: 21261981

1767. BMC Bioinformatics. 2011 Jan 18;12:23. doi: 10.1186/1471-2105-12-23.

easyDAS: automatic creation of DAS servers.

Gel Moreno B(1), Jenkinson AM, Jimenez RC, Messeguer Peypoch X, Hermjakob H.

Author information:

(1)Software Department, UPC-BarcelonaTech, Barcelona, Spain. bgel@lsi.upc.edu

BACKGROUND: The Distributed Annotation System (DAS) has proven to be a successful

way to publish and share biological data. Although there are more than 750 active

registered servers from around 50 organizations, setting up a DAS server

comprises a fair amount of work, making it difficult for many research groups to

share their biological annotations. Given the clear advantage that the

generalized sharing of relevant biological data is for the research community it

would be desirable to facilitate the sharing process.

RESULTS: Here we present easyDAS, a web-based system enabling anyone to publish

biological annotations with just some clicks. The system, available at

http://www.ebi.ac.uk/panda-srv/easydas is capable of reading different standard

data file formats, process the data and create a new publicly available DAS

source in a completely automated way. The created sources are hosted on the EBI

systems and can take advantage of its high storage capacity and network

connection, freeing the data provider from any network management work. easyDAS

is an open source project under the GNU LGPL license.

CONCLUSIONS: easyDAS is an automated DAS source creation system which can help

many researchers in sharing their biological data, potentially increasing the

amount of relevant biological data available to the scientific community.

DOI: 10.1186/1471-2105-12-23

PMCID: PMC3031199

PMID: 21244646 [Indexed for MEDLINE]

1768. Source Code Biol Med. 2011 Jan 17;6:3. doi: 10.1186/1751-0473-6-3.

WebGimm: An integrated web-based platform for cluster analysis, functional

analysis, and interactive visualization of results.

Joshi VK(1), Freudenberg JM, Hu Z, Medvedovic M.

Author information:

(1)Laboratory for Statistical Genomics and Systems Biology, Department of

Environmental Health, University of Cincinnati College of Medicine, 3223 Eden Av,

ML 56, Cincinnati OH 45267-0056, USA. medvedm@ucmail.uc.edu.

Cluster analysis methods have been extensively researched, but the adoption of

new methods is often hindered by technical barriers in their implementation and

use. WebGimm is a free cluster analysis web-service, and an open source general

purpose clustering web-server infrastructure designed to facilitate easy

deployment of integrated cluster analysis servers based on clustering and

functional annotation algorithms implemented in R. Integrated functional analyses

and interactive browsing of both, clustering structure and functional annotations

provides a complete analytical environment for cluster analysis and

interpretation of results. The Java Web Start client-based interface is modeled

after the familiar cluster/treeview packages making its use intuitive to a wide

array of biomedical researchers. For biomedical researchers, WebGimm provides an

avenue to access state of the art clustering procedures. For Bioinformatics

methods developers, WebGimm offers a convenient avenue to deploy their newly

developed clustering methods. WebGimm server, software and manuals can be freely

accessed at http://ClusterAnalysis.org/.

DOI: 10.1186/1751-0473-6-3

PMCID: PMC3033799

PMID: 21241501

1769. J Comput Chem. 2011 Jan 15;32(1):170-3. doi: 10.1002/jcc.21596.

NUPACK: Analysis and design of nucleic acid systems.

Zadeh JN(1), Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, Dirks RM,

Pierce NA.

Author information:

(1)Department of Bioengineering, California Institute of Technology, Pasadena,

California 91125, USA.

The Nucleic Acid Package (NUPACK) is a growing software suite for the analysis

and design of nucleic acid systems. The NUPACK web server (http://www.nupack.org)

currently enables:ANALYSIS: thermodynamic analysis of dilute solutions of

interacting nucleic acid strands.

DESIGN: sequence design for complexes of nucleic acid strands intended to adopt a

target secondary structure at equilibrium.Utilities: evaluation, display, and

annotation of equilibrium properties of a complex of nucleic acid strands. NUPACK

algorithms are formulated in terms of nucleic acid secondary structure. In most

cases, pseudoknots are excluded from the structural ensemble.

DOI: 10.1002/jcc.21596

PMID: 20645303 [Indexed for MEDLINE]

1770. Database (Oxford). 2011 Jan 6;2011:baq031. doi: 10.1093/database/baq031. Print

2011.

LPS-annotate: complete annotation of compositionally biased regions in the

protein knowledgebase.

Harbi D(1), Kumar M, Harrison PM.

Author information:

(1)Department of Biology, McGill University, Stewart Biology Building, 1205 Dr.

Penfield Ave., Montreal, QC, H3A 1B1, Canada.

Compositional bias (i.e. a skew in the composition of a biological sequence

towards a subset of residue types) can occur at a wide variety of scales, from

compositional biases of whole genomes, down to short regions in individual

protein and gene-DNA sequences that are compositionally biased (CB regions). Such

CB regions are made from a subset of residue types that are strewn along the

length of the region in an irregular way. Here, we have developed the database

server LPS-annotate, for the analysis of such CB regions, and protein disorder in

protein sequences. The algorithm defines compositional bias through a thorough

search for lowest-probability subsequences (LPSs) (i.e., the least likely

sequence regions in terms of composition). Users can (i) initially annotate CB

regions in input protein or nucleotide sequences of interest, and then (ii) query

a database of greater than 1,500,000 pre-calculated protein-CB regions, for

investigation of further functional hypotheses and inferences, about the specific

CB regions that were discovered, and their protein disorder propensities. We

demonstrate how a user can search for CB regions of similar compositional bias

and protein disorder, with a worked example. We show that our annotations

substantially augment the CB-region annotations that already exist in the UniProt

database, with more comprehensive annotation of more complex CB regions. Our

analysis indicates tens of thousands of CB regions that do not comprise globular

domains or transmembrane domains, and that do not have a propensity to protein

disorder, indicating a large cohort of protein-CB regions of biophysically

uncharacterized types. This server and database is a conceptually novel addition

to the workbench of tools now available to molecular biologists to generate

hypotheses and inferences about the proteins that they are investigating. It can

be accessed at http://libaio.biol.mcgill.ca/lps-annotate.html. Database URL:

http://libaio.biol.mcgill.ca/lps-annotate.html.

DOI: 10.1093/database/baq031

PMCID: PMC3017391

PMID: 21216786 [Indexed for MEDLINE]

1771. Source Code Biol Med. 2011 Jan 4;6(1):1. doi: 10.1186/1751-0473-6-1.

Tools for efficient epistasis detection in genome-wide association study.

Zhang X(1), Huang S, Zou F, Wang W.

Author information:

(1)Department of Computer Science, University of North Carolina at Chapel Hill,

Chapel Hill, NC, USA. xiang@cs.unc.edu.

BACKGROUND: Genome-wide association study (GWAS) aims to find genetic factors

underlying complex phenotypic traits, for which epistasis or gene-gene

interaction detection is often preferred over single-locus approach. However, the

computational burden has been a major hurdle to apply epistasis test in the

genome-wide scale due to a large number of single nucleotide polymorphism (SNP)

pairs to be tested.

RESULTS: We have developed a set of three efficient programs, FastANOVA, COE and

TEAM, that support epistasis test in a variety of problem settings in GWAS. These

programs utilize permutation test to properly control error rate such as

family-wise error rate (FWER) and false discovery rate (FDR). They guarantee to

find the optimal solutions, and significantly speed up the process of epistasis

detection in GWAS.

CONCLUSIONS: A web server with user interface and source codes are available at

the website http://www.csbio.unc.edu/epistasis/. The source codes are also

available at SourceForge http://sourceforge.net/projects/epistasis/.

DOI: 10.1186/1751-0473-6-1

PMCID: PMC3022563

PMID: 21205316

1772. Bioinformatics. 2011 Jan 1;27(1):137-9. doi: 10.1093/bioinformatics/btq594. Epub

2010 Nov 23.

Paintomics: a web based tool for the joint visualization of transcriptomics and

metabolomics data.

García-Alcalde F(1), García-López F, Dopazo J, Conesa A.

Author information:

(1)Bioinformatics and Genomics Department, Centro de Investigaciones Príncipe

Felipe, Valencia, Spain.

MOTIVATION: The development of the omics technologies such as transcriptomics,

proteomics and metabolomics has made possible the realization of systems biology

studies where biological systems are interrogated at different levels of

biochemical activity (gene expression, protein activity and/or metabolite

concentration). An effective approach to the analysis of these complex datasets

is the joined visualization of the disparate biomolecular data on the framework

of known biological pathways.

RESULTS: We have developed the Paintomics web server as an easy-to-use

bioinformatics resource that facilitates the integrated visual analysis of

experiments where transcriptomics and metabolomics data have been measured on

different conditions for the same samples. Basically, Paintomics takes complete

transcriptomics and metabolomics datasets, together with lists of significant

gene or metabolite changes, and paints this information on KEGG pathway maps.

AVAILABILITY: Paintomics is freely available at http://www.paintomics.org.

DOI: 10.1093/bioinformatics/btq594

PMCID: PMC3008637

PMID: 21098431 [Indexed for MEDLINE]

1773. Bioinformatics. 2011 Jan 1;27(1):132-3. doi: 10.1093/bioinformatics/btq610. Epub

2010 Nov 11.

Protein Peeling 3D: new tools for analyzing protein structures.

Gelly JC(1), de Brevern AG.

Author information:

(1)Dynamique des Structures et Interactions des Macromolécules Biologiques,

INSERM UMR-S 665, Université Paris Diderot-Paris 7, Paris, France.

jean-christophe.gelly@univ-paris-diderot.fr

We present an improved version of our Protein Peeling web server dedicated to the

analysis of protein structure architecture through the identification of protein

units produced by an iterative splitting algorithm. New features include

identification of structural domains, detection of unstructured terminal elements

and evaluation of the stability of protein unit structures.AVAILABILITY: The

website is free and open to all users with no login requirements at

http://www.dsimb.inserm.fr/dsimb-tools/peeling3.

DOI: 10.1093/bioinformatics/btq610

PMID: 21075745 [Indexed for MEDLINE]

1774. Bioinformatics. 2011 Jan 1;27(1):22-30. doi: 10.1093/bioinformatics/btq608. Epub

2010 Oct 28.

SPHINX--an algorithm for taxonomic binning of metagenomic sequences.

Mohammed MH(1), Ghosh TS, Singh NK, Mande SS.

Author information:

(1)Bio-sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services

Limited, Madhapur, Hyderabad, Andhra Pradesh, India.

MOTIVATION: Compared with composition-based binning algorithms, the binning

accuracy and specificity of alignment-based binning algorithms is significantly

higher. However, being alignment-based, the latter class of algorithms require

enormous amount of time and computing resources for binning huge metagenomic

datasets. The motivation was to develop a binning approach that can analyze

metagenomic datasets as rapidly as composition-based approaches, but nevertheless

has the accuracy and specificity of alignment-based algorithms. This article

describes a hybrid binning approach (SPHINX) that achieves high binning

efficiency by utilizing the principles of both 'composition'- and

'alignment'-based binning algorithms.

RESULTS: Validation results with simulated sequence datasets indicate that SPHINX

is able to analyze metagenomic sequences as rapidly as composition-based

algorithms. Furthermore, the binning efficiency (in terms of accuracy and

specificity of assignments) of SPHINX is observed to be comparable with results

obtained using alignment-based algorithms.

AVAILABILITY: A web server for the SPHINX algorithm is available at

http://metagenomics.atc.tcs.com/SPHINX/.

DOI: 10.1093/bioinformatics/btq608

PMID: 21030462 [Indexed for MEDLINE]

1775. Bioinformation. 2011;7(6):304-6. Epub 2011 Nov 20.

A protein short motif search tool using amino acid sequence and their secondary

structure assignment.

Venkataraman A, Chew TH, Hussein ZA, Shamsir MS.

We present the development of a web server, a protein short motif search tool

that allows users to simultaneously search for a protein sequence motif and its

secondary structure assignments. The web server is able to query very short

motifs searches against PDB structural data from the RCSB Protein Databank, with

the users defining the type of secondary structures of the amino acids in the

sequence motif. The output utilises 3D visualisation ability that highlights the

position of the motif in the structure and on the corresponding sequence.

Researchers can easily observe the locations and conformation of multiple motifs

among the results. Protein short motif search also has an application programming

interface (API) for interfacing with other bioinformatics tools.AVAILABILITY: The

database is available for free at http://birg3.fbb.utm.my/proteinsms.

PMCID: PMC3280500

PMID: 22355226

1776. Bioinformation. 2011;6(10):380-2. Epub 2011 Aug 2.

SSPred: A prediction server based on SVM for the identification and

classification of proteins involved in bacterial secretion systems.

Pundhir S(1), Kumar A.

Author information:

(1)School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore

(M.P) - 452001, India.

Protein secretion systems used by almost all bacteria are highly significant for

the normal existence and interaction of bacteria with their host. The

accumulation of genome sequence data in past few years has provided great

insights into the distribution and function of these secretion systems. In this

study, a support vector machine (SVM)- based method, SSPred was developed for the

automated functional annotation of proteins involved in secretion systems further

classifying them into five major sub-types (Type-I, Type-II, Type-III, Type-IV

and Sec systems). The dataset used in this study for training and testing was

obtained from KEGG and SwissProt database and was curated in order to avoid

redundancy. To overcome the problem of imbalance in positive and negative

dataset, an ensemble of SVM modules, each trained on a balanced subset of the

training data were used. Firstly, protein sequence features like amino-acid

composition (AAC), dipeptide composition (DPC) and physico-chemical composition

(PCC) were used to develop the SVM-based modules that achieved an average

accuracy of 84%, 85.17% and 82.59%, respectively. Secondly, a hybrid module

(hybrid-I) integrating all the previously used features was developed that

achieved an average accuracy of 86.12%. Another hybrid module (hybrid-II)

developed using evolutionary information of a protein sequence extracted from

position-specific scoring matrix and amino-acid composition achieved a maximum

average accuracy of 89.73%. On unbiased evaluation using an independent data set,

SSPred showed good prediction performance in identification and classification of

secretion systems. SSPred is a freely available World Wide Web server at

http//www.bioinformatics.org/sspred.

PMCID: PMC3163916

PMID: 21904425

1777. Bioinformation. 2011;6(7):288-90. Epub 2011 Jun 23.

OntoVisT: A general purpose Ontological Visualization Tool.

Srivastava AK(1), Sahni NS.

Author information:

(1)School of Computational and Integrative Sciences, Jawaharlal Nehru University,

New Delhi - 110067, India.

Ontologies have emerged as a fast growing research topic in the area of semantic

web during last decade. Currently there are 204 ontologies that are available

through OBO Foundry and BioPortal. Several excellent tools for navigating the

ontological structure are available, however most of them are dedicated to a

specific annotation data or integrated with specific analysis applications, and

do not offer flexibility in terms of general-purpose usage for ontology

exploration. We developed OntoVisT, a web based ontological visualization tool.

This application is designed for interactive visualization of any ontological

hierarchy for a specific node of interest, up to the chosen level of children

and/or ancestor. It takes any ontology file in OBO format as input and generates

output as DAG hierarchical graph for the chosen query. To enhance the navigation

capabilities of complex networks, we have embedded several features such as

search criteria, zoom in/out, center focus, nearest neighbor highlights and mouse

hover events. The application has been tested on all 72 data sets available in

OBO format through OBO foundry. The results for few of them can be accessed

through OntoVisT-Gallery.AVAILABILITY: The database is available for free at

http://ccbb.jnu.ac.in/OntoVisT.html.

PMCID: PMC3124697

PMID: 21738333

1778. BMC Bioinformatics. 2011;12 Suppl 13:S9. doi: 10.1186/1471-2105-12-S13-S9. Epub

2011 Nov 30.

HabiSign: a novel approach for comparison of metagenomes and rapid identification

of habitat-specific sequences.

Ghosh TS(1), Mohammed MH, Rajasingh H, Chadaram S, Mande SS.

Author information:

(1)Bio-sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services

Limited, 1 Software Units Layout, Madhapur, Hyderabad - 500081, Andhra Pradesh,

India.

BACKGROUND: One of the primary goals of comparative metagenomic projects is to

study the differences in the microbial communities residing in diverse

environments. Besides providing valuable insights into the inherent structure of

the microbial populations, these studies have potential applications in several

important areas of medical research like disease diagnostics, detection of

pathogenic contamination and identification of hitherto unknown pathogens. Here

we present a novel and rapid, alignment-free method called HabiSign, which

utilizes patterns of tetra-nucleotide usage in microbial genomes to bring out the

differences in the composition of both diverse and related microbial communities.

RESULTS: Validation results show that the metagenomic signatures obtained using

the HabiSign method are able to accurately cluster metagenomes at biome,

phenotypic and species levels, as compared to an average tetranucleotide

frequency based approach and the recently published dinucleotide relative

abundance based approach. More importantly, the method is able to identify

subsets of sequences that are specific to a particular habitat. Apart from this,

being alignment-free, the method can rapidly compare and group multiple

metagenomic data sets in a short span of time.

CONCLUSIONS: The proposed method is expected to have immense applicability in

diverse areas of metagenomic research ranging from disease diagnostics and

pathogen detection to bio-prospecting. A web-server for the HabiSign algorithm is

available at http://metagenomics.atc.tcs.com/HabiSign/.

DOI: 10.1186/1471-2105-12-S13-S9

PMCID: PMC3278849

PMID: 22373355 [Indexed for MEDLINE]

1779. BMC Bioinformatics. 2011;12 Suppl 13:S5. doi: 10.1186/1471-2105-12-S13-S5. Epub

2011 Nov 30.

Prediction of dinucleotide-specific RNA-binding sites in proteins.

Fernandez M(1), Kumagai Y, Standley DM, Sarai A, Mizuguchi K, Ahmad S.

Author information:

(1)Kyushu Institute of Technology, Fukuoka, Japan.

BACKGROUND: Regulation of gene expression, protein synthesis, replication and

assembly of many viruses involve RNA-protein interactions. Although some

successful computational tools have been reported to recognize RNA binding sites

in proteins, the problem of specificity remains poorly investigated. After the

nucleotide base composition, the dinucleotide is the smallest unit of RNA

sequence information and many RNA-binding proteins simply bind to regions

enriched in one dinucleotide. Interaction preferences of protein subsequences and

dinucleotides can be inferred from protein-RNA complex structures, enabling a

training-based prediction approach.

RESULTS: We analyzed basic statistics of amino acid-dinucleotide contacts in

protein-RNA complexes and found their pairing preferences could be identified.

Using a standard approach to represent protein subsequences by their evolutionary

profile, we trained neural networks to predict multiclass target vectors

corresponding to 16 possible contacting dinucleotide subsequences. In the

cross-validation experiments, the accuracies of the optimum network, measured as

areas under the curve (AUC) of the receiver operating characteristic (ROC)

graphs, were in the range of 65-80%.

CONCLUSIONS: Dinucleotide-specific contact predictions have also been extended to

the prediction of interacting protein and RNA fragment pairs, which shows the

applicability of this method to predict targets of RNA-binding proteins. A web

server predicting the 16-dimensional contact probability matrix directly from a

user-defined protein sequence was implemented and made available at:

http://tardis.nibio.go.jp/netasa/srcpred.

DOI: 10.1186/1471-2105-12-S13-S5

PMCID: PMC3278845

PMID: 22373260 [Indexed for MEDLINE]

1780. BMC Bioinformatics. 2011;12 Suppl 13:S4. doi: 10.1186/1471-2105-12-S13-S4. Epub

2011 Nov 30.

PTIGS-IdIt, a system for species identification by DNA sequences of the psbA-trnH

intergenic spacer region.

Liu C(1), Liang D, Gao T, Pang X, Song J, Yao H, Han J, Liu Z, Guan X, Jiang K,

Li H, Chen S.

Author information:

(1)The Key Laboratory of Bioactive Substances and Resources Utilization of

Chinese Herbal Medicine, Ministry of Education, Institute of Medicinal Plant

Development, Chinese Academy of Medical Sciences & Peking Union Medical College,

Beijing 100193, P R China. cliu@implad.ac.cn

BACKGROUND: DNA barcoding technology, which uses a short piece of DNA sequence to

identify species, has wide ranges of applications. Until today, a universal DNA

barcode marker for plants remains elusive. The rbcL and matK regions have been

proposed as the "core barcode" for plants and the ITS2 and psbA-trnH intergenic

spacer (PTIGS) regions were later added as supplemental barcodes. The use of

PTIGS region as a supplemental barcode has been limited by the lack of

computational tools that can handle significant insertions and deletions in the

PTIGS sequences. Here, we compared the most commonly used alignment-based and

alignment-free methods and developed a web server to allow the biologists to

carry out PTIGS-based DNA barcoding analyses.

RESULTS: First, we compared several alignment-based methods such as BLAST and

those calculating P distance and Edit distance, alignment-free methods

Di-Nucleotide Frequency Profile (DNFP) and their combinations. We found that the

DNFP and Edit-distance methods increased the identification success rate to ~80%,

20% higher than the most commonly used BLAST method. Second, the combined methods

showed overall better success rate and performance. Last, we have developed a web

server that allows (1) retrieving various sub-regions and the consensus sequences

of PTIGS, (2) annotating novel PTIGS sequences, (3) determining species identity

by PTIGS sequences using eight methods, and (4) examining identification

efficiency and performance of the eight methods for various taxonomy groups.

CONCLUSIONS: The Edit distance and the DNFP methods have the highest

discrimination powers. Hybrid methods can be used to achieve significant

improvement in performance. These methods can be extended to applications using

the core barcodes and the other supplemental DNA barcode ITS2. To our knowledge,

the web server developed here is the only one that allows species determination

based on PTIGS sequences. The web server can be accessed at

http://psba-trnh-plantidit.dnsalias.org.

DOI: 10.1186/1471-2105-12-S13-S4

PMCID: PMC3278844

PMID: 22373238 [Indexed for MEDLINE]

1781. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:3221-4. doi:

10.1109/IEMBS.2011.6090876.

Structure-based prediction of protein activity changes: assessing the impact of

single residue replacements.

Masso M(1), Vaisman II.

Author information:

(1)Laboratory for Structural Bioinformatics, School of Systems Biology, George

Mason University, Manassas, VA 20110, USA. mmasso@gmu.edu

A computational mutagenesis methodology founded upon a structure-dependent and

knowledge-based four-body statistical potential is utilized in generating feature

vectors that characterize over 8500 individual amino acid substitutions occurring

in seven proteins, each mutant having been experimentally ascertained for its

relative effect on native protein activity. The proteins are diverse with respect

to host organism (viral, bacterial, human) and function (enzymatic, nucleic acid

binding, signaling), the structures span all four major SCOP classifications, and

the mutations occur at positions well distributed throughout the seven

structures. Implementation of the random forest algorithm, for classifying mutant

activity as either unaffected or affected relative to the native protein, yields

84% accuracy based on tenfold cross-validation. A freely available online server

for obtaining predictions with the trained model, which also displays 84%

accuracy on an independent test set of mutants, is available at

http://proteins.gmu.edu/automute/AUTO-MUTE\_Activity.html.

DOI: 10.1109/IEMBS.2011.6090876

PMID: 22255025 [Indexed for MEDLINE]

1782. Genome Biol Evol. 2011;3:1265-75. doi: 10.1093/gbe/evr101. Epub 2011 Oct 4.

Inference of gain and loss events from phyletic patterns using stochastic mapping

and maximum parsimony--a simulation study.

Cohen O(1), Pupko T.

Author information:

(1)Department of Cell Research and Immunology, Tel Aviv University, Tel Aviv,

Israel.

Bacterial evolution is characterized by frequent gain and loss events of gene

families. These events can be inferred from phyletic pattern data-a compact

representation of gene family repertoire across multiple genomes. The maximum

parsimony paradigm is a classical and prevalent approach for the detection of

gene family gains and losses mapped on specific branches. We and others have

previously developed probabilistic models that aim to account for the gain and

loss stochastic dynamics. These models are a critical component of a methodology

termed stochastic mapping, in which probabilities and expectations of gain and

loss events are estimated for each branch of an underlying phylogenetic tree. In

this work, we present a phyletic pattern simulator in which the gain and loss

dynamics are assumed to follow a continuous-time Markov chain along the tree.

Various models and options are implemented to make the simulation software useful

for a large number of studies in which binary (presence/absence) data are

analyzed. Using this simulation software, we compared the ability of the maximum

parsimony and the stochastic mapping approaches to accurately detect gain and

loss events along the tree. Our simulations cover a large array of evolutionary

scenarios in terms of the propensities for gene family gains and losses and the

variability of these propensities among gene families. Although in all simulation

schemes, both methods obtain relatively low levels of false positive rates,

stochastic mapping outperforms maximum parsimony in terms of true positive rates.

We further studied the factors that influence the performance of both methods. We

find, for example, that the accuracy of maximum parsimony inference is

substantially reduced when the goal is to map gain and loss events along internal

branches of the phylogenetic tree. Furthermore, the accuracy of stochastic

mapping is reduced with smaller data sets (limited number of gene families) due

to unreliable estimation of branch lengths. Our simulator and simulation results

are additionally relevant for the analysis of other types of binary-coded data,

such as the existence of homologues restriction sites, gaps, and introns, to name

a few. Both the simulation software and the inference methodology are freely

available at a user-friendly server: http://gloome.tau.ac.il/.

DOI: 10.1093/gbe/evr101

PMCID: PMC3215202

PMID: 21971516 [Indexed for MEDLINE]

1783. Genome Inform. 2011;25(1):1-11.

Database for crude drugs and Kampo medicine.

Arita M(1), Yoshimoto M, Suwa K, Hirai A, Kanaya S, Shibahara N, Tanaka K.

Author information:

(1)Department of Biophysics and Biochemistry, Graduate School of Science, The

University of Tokyo, Bunkyo-ku, Tokyo, Japan. arita@bi.s.u-tokyo.ac.jp

A wiki-based repository for crude drugs and Kampo medicine is introduced. It

provides taxonomic and chemical information for 158 crude drugs and 348

prescriptions of the traditional Kampo medicine in Japan, which is a variation of

ancient Chinese medicine. The system is built on MediaWiki with extensions for

inline page search and for sending user-input elements to the server. These

functions together realize implementation of word checks and data integration at

the user-level. In this scheme, any user can participate in creating an

integrated database with controlled vocabularies on the wiki system. Our

implementation and data are accessible at http://metabolomics.jp/wiki/.

PMID: 22230935 [Indexed for MEDLINE]

1784. Hum Mutat. 2011 Jan;32(1):E1948-58. doi: 10.1002/humu.21393.

PRO-MINE: A bioinformatics repository and analytical tool for TARDBP mutations.

Pinto S(1), Vlahoviček K, Buratti E.

Author information:

(1)Department of Molecular Biology, University of Zagreb, Croatia.

TDP-43 is a multifunctional RNA-binding protein found to be a major protein

component of intracellular inclusions found in neurodegenerative disorders such

as Fronto Temporal Lobar Degeneration, Amyotrophic Lateral Sclerosis, and

Alzheimer Disease. PRO-MINE (PROtein Mutations In NEurodegeneration) is a

database populated with manually curated data from the literature regarding all

TDP-43/TDP43/TARDBP gene disease-associated mutations identified to date. A web

server interface has been developed to query the database and to provide tools

for the analysis of already reported or novel TDP-43 gene mutations. As is

usually the case with genetic association studies, assessing the potential impact

of identified mutations is of crucial importance, and in order to avoid

prediction biases it is essential to compare the prediction results. However, in

most cases mutations have to be submitted separately to various prediction tools

and the individual results manually merged together afterwards. The implemented

web server aims to overcome the problem by providing simultaneous access to

several prediction tools and by displaying the results into a single output.

Furthermore, the results are displayed together in a comprehensive output for a

more convenient analysis and are enriched with additional information about

mutations. In addition, our web server can also display the mutation(s) of

interest within an alignment of annotated TDP-43 protein sequences from different

vertebrate species. In this way, the degree of sequence conservation where the

mutation(s) occur can be easily tracked and visualized. The web server is freely

available to researchers and can be accessed at http://bioinfo.hr/pro-mine.

© 2010 Wiley-Liss, Inc.

DOI: 10.1002/humu.21393

PMCID: PMC3038324

PMID: 21031599 [Indexed for MEDLINE]

1785. Infect Immun. 2011 Jan;79(1):23-32. doi: 10.1128/IAI.00537-10. Epub 2010 Oct 25.

Computational prediction of type III and IV secreted effectors in gram-negative

bacteria.

McDermott JE(1), Corrigan A, Peterson E, Oehmen C, Niemann G, Cambronne ED, Sharp

D, Adkins JN, Samudrala R, Heffron F.

Author information:

(1)Computational Biology and Bioinformatics Group, Pacific Northwest National

Laboratory, MSIN: J4-33, 902 Battelle Boulevard, P.O. Box 999, Richland, WA

99352, USA. Jason.McDermott@pnl.gov

In this review, we provide an overview of the methods employed in four recent

studies that described novel methods for computational prediction of secreted

effectors from type III and IV secretion systems in Gram-negative bacteria. We

present the results of these studies in terms of performance at accurately

predicting secreted effectors and similarities found between secretion signals

that may reflect biologically relevant features for recognition. We discuss the

Web-based tools for secreted effector prediction described in these studies and

announce the availability of our tool, the SIEVE server

(http://www.sysbep.org/sieve). Finally, we assess the accuracies of the three

type III effector prediction methods on a small set of proteins not known prior

to the development of these tools that we recently discovered and validated using

both experimental and computational approaches. Our comparison shows that all

methods use similar approaches and, in general, arrive at similar conclusions. We

discuss the possibility of an order-dependent motif in the secretion signal,

which was a point of disagreement in the studies. Our results show that there may

be classes of effectors in which the signal has a loosely defined motif and

others in which secretion is dependent only on compositional biases.

Computational prediction of secreted effectors from protein sequences represents

an important step toward better understanding the interaction between pathogens

and hosts.

DOI: 10.1128/IAI.00537-10

PMCID: PMC3019878

PMID: 20974833 [Indexed for MEDLINE]

1786. Int J Data Min Bioinform. 2011;5(1):38-51.

Understandable learning machine system design for transmembrane or embedded

membrane segments prediction.

Hu HJ(1), Harrison RW, Tai PC, Pan Y.

Author information:

(1)Department of Computer Science, Georgia State University, Atlanta, GA 30303,

USA. haejin.hu@gmail.com

We modified an existing association rule-based classifier CPAR to improve

traditional black box model based learning machine approaches on Transmembrane

(TM) segment prediction. The modified classifier was improved further by

combining with SVM. The experimental results indicate that this hybrid scheme

offers biologically meaningful rules on TM/EM segment prediction while

maintaining the performance almost as well as the SVM method. The evaluation of

the sturdiness and the Receiver Operating Characteristic (ROC) curve analysis

proved that this new scheme is robust and competent with SVM on TM/EM segment

prediction. The prediction server is available at http://bmcc2.cs.gsu.edu/

haeh2/.

PMID: 21491843 [Indexed for MEDLINE]

1787. J Biomed Biotechnol. 2011;2011. pii: 839862. doi: 10.1155/2011/839862. Epub 2010

Sep 5.

metaP-server: a web-based metabolomics data analysis tool.

Kastenmüller G(1), Römisch-Margl W, Wägele B, Altmaier E, Suhre K.

Author information:

(1)Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München,

German Research Center for Environmental Health, Ingolstädter Landstraße 1,

Neuherberg, Germany.

Metabolomics is an emerging field that is based on the quantitative measurement

of as many small organic molecules occurring in a biological sample as possible.

Due to recent technical advances, metabolomics can now be used widely as an

analytical high-throughput technology in drug testing and epidemiological

metabolome and genome wide association studies. Analogous to chip-based gene

expression analyses, the enormous amount of data produced by modern kit-based

metabolomics experiments poses new challenges regarding their biological

interpretation in the context of various sample phenotypes. We developed

metaP-server to facilitate data interpretation. metaP-server provides automated

and standardized data analysis for quantitative metabolomics data, covering the

following steps from data acquisition to biological interpretation: (i) data

quality checks, (ii) estimation of reproducibility and batch effects, (iii)

hypothesis tests for multiple categorical phenotypes, (iv) correlation tests for

metric phenotypes, (v) optionally including all possible pairs of metabolite

concentration ratios, (vi) principal component analysis (PCA), and (vii) mapping

of metabolites onto colored KEGG pathway maps. Graphical output is clickable and

cross-linked to sample and metabolite identifiers. Interactive coloring of PCA

and bar plots by phenotype facilitates on-line data exploration. For users of

commercial metabolomics kits, cross-references to the HMDB, LipidMaps, KEGG,

PubChem, and CAS databases are provided. metaP-server is freely accessible at

http://metabolomics.helmholtz-muenchen.de/metap2/.

DOI: 10.1155/2011/839862

PMCID: PMC2946609

PMID: 20936179 [Indexed for MEDLINE]

1788. Methods Enzymol. 2011;498:173-88. doi: 10.1016/B978-0-12-385120-8.00008-5.

A step-by-step introduction to rule-based design of synthetic genetic constructs

using GenoCAD.

Wilson ML(1), Hertzberg R, Adam L, Peccoud J.

Author information:

(1)Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA.

GenoCAD is an open source web-based system that provides a streamlined,

rule-driven process for designing genetic sequences. GenoCAD provides a graphical

interface that allows users to design sequences consistent with formalized design

strategies specific to a domain, organization, or project. Design strategies

include limited sets of user-defined parts and rules indicating how these parts

are to be combined in genetic constructs. In addition to reducing design time to

minutes, GenoCAD improves the quality and reliability of the finished sequence by

ensuring that the designs follow established rules of sequence construction.

GenoCAD.org is a publicly available instance of GenoCAD that can be found at

www.genocad.org. The source code and latest build are available from SourceForge

to allow advanced users to install and customize GenoCAD for their unique needs.

This chapter focuses primarily on how the GenoCAD tools can be used to organize

genetic parts into customized personal libraries, then how these libraries can be

used to design sequences. In addition, GenoCAD's parts management system and

search capabilities are described in detail. Instructions are provided for

installing a local instance of GenoCAD on a server. Some of the future

enhancements of this rapidly evolving suite of applications are briefly

described.

Copyright © 2011 Elsevier Inc. All rights reserved.

DOI: 10.1016/B978-0-12-385120-8.00008-5

PMID: 21601678 [Indexed for MEDLINE]

1789. Nucleic Acids Res. 2011 Jan;39(Database issue):D830-4. doi: 10.1093/nar/gkq1235.

Epub 2010 Dec 1.

Bovine Genome Database: integrated tools for genome annotation and discovery.

Childers CP(1), Reese JT, Sundaram JP, Vile DC, Dickens CM, Childs KL, Salih H,

Bennett AK, Hagen DE, Adelson DL, Elsik CG.

Author information:

(1)Department of Biology, Georgetown University, Washington, DC 20057, USA.

The Bovine Genome Database (BGD; http://BovineGenome.org) strives to improve

annotation of the bovine genome and to integrate the genome sequence with other

genomics data. BGD includes GBrowse genome browsers, the Apollo Annotation

Editor, a quantitative trait loci (QTL) viewer, BLAST databases and gene pages.

Genome browsers, available for both scaffold and chromosome coordinate systems,

display the bovine Official Gene Set (OGS), RefSeq and Ensembl gene models,

non-coding RNA, repeats, pseudogenes, single-nucleotide polymorphism, markers,

QTL and alignments to complementary DNAs, ESTs and protein homologs. The Bovine

QTL viewer is connected to the BGD Chromosome GBrowse, allowing for the

identification of candidate genes underlying QTL. The Apollo Annotation Editor

connects directly to the BGD Chado database to provide researchers with remote

access to gene evidence in a graphical interface that allows editing and creating

new gene models. Researchers may upload their annotations to the BGD server for

review and integration into the subsequent release of the OGS. Gene pages display

information for individual OGS gene models, including gene structure, transcript

variants, functional descriptions, gene symbols, Gene Ontology terms, annotator

comments and links to National Center for Biotechnology Information and Ensembl.

Each gene page is linked to a wiki page to allow input from the research

community.

DOI: 10.1093/nar/gkq1235

PMCID: PMC3013744

PMID: 21123190 [Indexed for MEDLINE]

1790. Nucleic Acids Res. 2011 Jan;39(Database issue):D220-4. doi: 10.1093/nar/gkq1157.

Epub 2010 Nov 24.

MIPS: curated databases and comprehensive secondary data resources in 2010.

Mewes HW(1), Ruepp A, Theis F, Rattei T, Walter M, Frishman D, Suhre K, Spannagl

M, Mayer KF, Stümpflen V, Antonov A.

Author information:

(1)Institute for Bioinformatics and Systems Biology, MIPS, Helmholtz Center F

Health and Environment, Ingolstädter Landstr 1, D-85764 Neuherberg, Germany.

w.mewes@helmholtz-muenchen.de

The Munich Information Center for Protein Sequences (MIPS at the Helmholtz Center

for Environmental Health, Neuherberg, Germany) has many years of experience in

providing annotated collections of biological data. Selected data sets of high

relevance, such as model genomes, are subjected to careful manual curation, while

the bulk of high-throughput data is annotated by automatic means. High-quality

reference resources developed in the past and still actively maintained include

Saccharomyces cerevisiae, Neurospora crassa and Arabidopsis thaliana genome

databases as well as several protein interaction data sets (MPACT, MPPI and

CORUM). More recent projects are PhenomiR, the database on microRNA-related

phenotypes, and MIPS PlantsDB for integrative and comparative plant genome

research. The interlinked resources SIMAP and PEDANT provide homology

relationships as well as up-to-date and consistent annotation for 38,000,000

protein sequences. PPLIPS and CCancer are versatile tools for proteomics and

functional genomics interfacing to a database of compilations from gene lists

extracted from literature. A novel literature-mining tool, EXCERBT, gives access

to structured information on classified relations between genes, proteins,

phenotypes and diseases extracted from Medline abstracts by semantic analysis.

All databases described here, as well as the detailed descriptions of our

projects can be accessed through the MIPS WWW server

(http://mips.helmholtz-muenchen.de).

DOI: 10.1093/nar/gkq1157

PMCID: PMC3013725

PMID: 21109531 [Indexed for MEDLINE]

1791. Nucleic Acids Res. 2011 Jan;39(Database issue):D465-74. doi: 10.1093/nar/gkq1091.

Epub 2010 Nov 19.

ModBase, a database of annotated comparative protein structure models, and

associated resources.

Pieper U(1), Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H,

Yang Z, Meng EC, Pettersen EF, Huang CC, Datta RS, Sampathkumar P, Madhusudhan

MS, Sjölander K, Ferrin TE, Burley SK, Sali A.

Author information:

(1)Department of Bioengineering and Therapeutic Sciences, Department of

Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences,

University of California at San Francisco, CA 94158, USA.

ModBase (http://salilab.org/modbase) is a database of annotated comparative

protein structure models. The models are calculated by ModPipe, an automated

modeling pipeline that relies primarily on Modeller for fold assignment,

sequence-structure alignment, model building and model assessment

(http://salilab.org/modeller/). ModBase currently contains 10,355,444 reliable

models for domains in 2,421,920 unique protein sequences. ModBase allows users to

update comparative models on demand, and request modeling of additional sequences

through an interface to the ModWeb modeling server (http://salilab.org/modweb).

ModBase models are available through the ModBase interface as well as the Protein

Model Portal (http://www.proteinmodelportal.org/). Recently developed associated

resources include the SALIGN server for multiple sequence and structure alignment

(http://salilab.org/salign), the ModEval server for predicting the accuracy of

protein structure models (http://salilab.org/modeval), the PCSS server for

predicting which peptides bind to a given protein (http://salilab.org/pcss) and

the FoXS server for calculating and fitting Small Angle X-ray Scattering profiles

(http://salilab.org/foxs).

DOI: 10.1093/nar/gkq1091

PMCID: PMC3013688

PMID: 21097780 [Indexed for MEDLINE]

1792. Nucleic Acids Res. 2011 Jan;39(Database issue):D420-6. doi: 10.1093/nar/gkq1001.

Epub 2010 Nov 19.

Extending CATH: increasing coverage of the protein structure universe and linking

structure with function.

Cuff AL(1), Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N,

Pellegrini-Calace M, Jones D, Thornton J, Orengo CA.

Author information:

(1)Institute of Structural and Molecular Biology, University College London,

Darwin Building, Gower Street, London WC1E 6BT, UK. cuff@biochem.ucl.ac.uk

CATH version 3.3 (class, architecture, topology, homology) contains 128,688

domains, 2386 homologous superfamilies and 1233 fold groups, and reflects a major

focus on classifying structural genomics (SG) structures and transmembrane

proteins, both of which are likely to add structural novelty to the database and

therefore increase the coverage of protein fold space within CATH. For CATH

version 3.4 we have significantly improved the presentation of sequence

information and associated functional information for CATH superfamilies. The

CATH superfamily pages now reflect both the functional and structural diversity

within the superfamily and include structural alignments of close and distant

relatives within the superfamily, annotated with functional information and

details of conserved residues. A significantly more efficient search function for

CATH has been established by implementing the search server Solr

(http://lucene.apache.org/solr/). The CATH v3.4 webpages have been built using

the Catalyst web framework.

DOI: 10.1093/nar/gkq1001

PMCID: PMC3013636

PMID: 21097779 [Indexed for MEDLINE]

1793. Nucleic Acids Res. 2011 Jan;39(Database issue):D1085-94. doi:

10.1093/nar/gkq1148. Epub 2010 Nov 13.

Gramene database in 2010: updates and extensions.

Youens-Clark K(1), Buckler E, Casstevens T, Chen C, Declerck G, Derwent P,

Dharmawardhana P, Jaiswal P, Kersey P, Karthikeyan AS, Lu J, McCouch SR, Ren L,

Spooner W, Stein JC, Thomason J, Wei S, Ware D.

Author information:

(1)Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724,

USA. kclark@cshl.edu

Now in its 10th year, the Gramene database (http://www.gramene.org) has grown

from its primary focus on rice, the first fully-sequenced grass genome, to become

a resource for major model and crop plants including Arabidopsis, Brachypodium,

maize, sorghum, poplar and grape in addition to several species of rice. Gramene

began with the addition of an Ensembl genome browser and has expanded in the last

decade to become a robust resource for plant genomics hosting a wide array of

data sets including quantitative trait loci (QTL), metabolic pathways, genetic

diversity, genes, proteins, germplasm, literature, ontologies and a

fully-structured markers and sequences database integrated with genome browsers

and maps from various published studies (genetic, physical, bin, etc.). In

addition, Gramene now hosts a variety of web services including a Distributed

Annotation Server (DAS), BLAST and a public MySQL database. Twice a year, Gramene

releases a major build of the database and makes interim releases to correct

errors or to make important updates to software and/or data.

DOI: 10.1093/nar/gkq1148

PMCID: PMC3013721

PMID: 21076153 [Indexed for MEDLINE]

1794. Nucleic Acids Res. 2011 Jan;39(Database issue):D1118-22. doi:

10.1093/nar/gkq1120. Epub 2010 Nov 8.

AGRIS: the Arabidopsis Gene Regulatory Information Server, an update.

Yilmaz A(1), Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E.

Author information:

(1)Department of Plant Cellular and Molecular Biology and Plant Biotechnology

Center, The Ohio State University, Columbus, OH, USA. yilmaz.11@osu.edu

The Arabidopsis Gene Regulatory Information Server (AGRIS;

http://arabidopsis.med.ohio-state.edu/) provides a comprehensive resource for

gene regulatory studies in the model plant Arabidopsis thaliana. Three

interlinked databases, AtTFDB, AtcisDB and AtRegNet, furnish comprehensive and

updated information on transcription factors (TFs), predicted and experimentally

verified cis-regulatory elements (CREs) and their interactions, respectively. In

addition to significant contributions in the identification of the entire set of

TF-DNA interactions, which are the key to understand the gene regulatory networks

that govern Arabidopsis gene expression, tools recently incorporated into AGRIS

include the complete set of words length 5-15 present in the Arabidopsis genome

and the integration of AtRegNet with visualization tools, such as the recently

developed ReIN application. All the information in AGRIS is publicly available

and downloadable upon registration.

DOI: 10.1093/nar/gkq1120

PMCID: PMC3013708

PMID: 21059685 [Indexed for MEDLINE]

1795. Nucleic Acids Res. 2011 Jan;39(Database issue):D975-9. doi: 10.1093/nar/gkq1024.

Epub 2010 Nov 2.

CCDB: a curated database of genes involved in cervix cancer.

Agarwal SM(1), Raghav D, Singh H, Raghava GP.

Author information:

(1)Bioinformatics Division, Institute of Cytology and Preventive Oncology, I-7,

Sector 39, Noida 201301, India. smagarwal@yahoo.com

The Cervical Cancer gene DataBase (CCDB, http://crdd.osdd.net/raghava/ccdb) is a

manually curated catalog of experimentally validated genes that are thought, or

are known to be involved in the different stages of cervical carcinogenesis. In

spite of the large women population that is presently affected from this

malignancy still at present, no database exists that catalogs information on

genes associated with cervical cancer. Therefore, we have compiled 537 genes in

CCDB that are linked with cervical cancer causation processes such as

methylation, gene amplification, mutation, polymorphism and change in expression

level, as evident from published literature. Each record contains details related

to gene like architecture (exon-intron structure), location, function, sequences

(mRNA/CDS/protein), ontology, interacting partners, homology to other eukaryotic

genomes, structure and links to other public databases, thus augmenting CCDB with

external data. Also, manually curated literature references have been provided to

support the inclusion of the gene in the database and establish its association

with cervix cancer. In addition, CCDB provides information on microRNA altered in

cervical cancer as well as search facility for querying, several browse options

and an online tool for sequence similarity search, thereby providing researchers

with easy access to the latest information on genes involved in cervix cancer.

DOI: 10.1093/nar/gkq1024

PMCID: PMC3013652

PMID: 21045064 [Indexed for MEDLINE]

1796. Nucleic Acids Res. 2011 Jan;39(Database issue):D70-4. doi: 10.1093/nar/gkq1061.

Epub 2010 Oct 29.

The Gypsy Database (GyDB) of mobile genetic elements: release 2.0.

Llorens C(1), Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D,

Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, Maumus F,

Munoz-Pomer A, Sempere JM, Latorre A, Moya A.

Author information:

(1)Biotechvana, Parc Científic, Universitat de València, Calle Catedrático José

Beltrán 2, 46980 Paterna, València, Spain. carlos.llorens@uv.es

This article introduces the second release of the Gypsy Database of Mobile

Genetic Elements (GyDB 2.0): a research project devoted to the evolutionary

dynamics of viruses and transposable elements based on their phylogenetic

classification (per lineage and protein domain). The Gypsy Database (GyDB) is a

long-term project that is continuously progressing, and that owing to the high

molecular diversity of mobile elements requires to be completed in several

stages. GyDB 2.0 has been powered with a wiki to allow other researchers

participate in the project. The current database stage and scope are long

terminal repeats (LTR) retroelements and relatives. GyDB 2.0 is an update based

on the analysis of Ty3/Gypsy, Retroviridae, Ty1/Copia and Bel/Pao LTR

retroelements and the Caulimoviridae pararetroviruses of plants. Among other

features, in terms of the aforementioned topics, this update adds: (i) a variety

of descriptions and reviews distributed in multiple web pages; (ii) protein-based

phylogenies, where phylogenetic levels are assigned to distinct classified

elements; (iii) a collection of multiple alignments, lineage-specific hidden

Markov models and consensus sequences, called GyDB collection; (iv) updated

RefSeq databases and BLAST and HMM servers to facilitate sequence

characterization of new LTR retroelement and caulimovirus queries; and (v) a

bibliographic server. GyDB 2.0 is available at http://gydb.org.

DOI: 10.1093/nar/gkq1061

PMCID: PMC3013669

PMID: 21036865 [Indexed for MEDLINE]

1797. Nucleic Acids Res. 2011 Jan;39(Database issue):D718-23. doi: 10.1093/nar/gkq962.

Epub 2010 Oct 21.

3did: identification and classification of domain-based interactions of known

three-dimensional structure.

Stein A(1), Céol A, Aloy P.

Author information:

(1)Institute for Research in Biomedicine, Join IRB-BSC Program in Computational

Biology, 08028 Barcelona, Spain.

The database of three-dimensional interacting domains (3did) is a collection of

protein interactions for which high-resolution three-dimensional structures are

known. 3did exploits the availability of structural data to provide molecular

details on interactions between two globular domains as well as novel

domain-peptide interactions, derived using a recently published method from our

lab. The interface residues are presented for each interaction type individually,

plus global domain interfaces at which one or more partners (domains or peptides)

bind. The 3did web server at http://3did.irbbarcelona.org visualizes these

interfaces along with atomic details of individual interactions using Jmol. The

complete contents are also available for download.

DOI: 10.1093/nar/gkq962

PMCID: PMC3013799

PMID: 20965963 [Indexed for MEDLINE]

1798. Nucleic Acids Res. 2011 Jan;39(Database issue):D632-6. doi: 10.1093/nar/gkq918.

Epub 2010 Oct 14.

ParameciumDB in 2011: new tools and new data for functional and comparative

genomics of the model ciliate Paramecium tetraurelia.

Arnaiz O(1), Sperling L.

Author information:

(1)Centre de Génétique Moléculaire, CNRS FRE3144, Avenue de la Terrasse, 91198

Gif-sur-Yvette, France.

ParameciumDB is a community model organism database built with the GMOD toolkit

to integrate the genome and biology of the ciliate Paramecium tetraurelia. Over

the last four years, post-genomic data from proteome and transcriptome studies

has been incorporated along with predicted orthologs in 33 species, annotations

from the community and publications from the scientific literature. Available

tools include BioMart for complex queries, GBrowse2 for genome browsing, the

Apollo genome editor for expert curation of gene models, a Blast server, a motif

finder, and a wiki for protocols, nomenclature guidelines and other

documentation. In-house tools have been developed for ontology browsing and

evaluation of off-target RNAi matches. Now ready for next-generation deep

sequencing data and the genomes of other Paramecium species, this open-access

resource is available at http://paramecium.cgm.cnrs-gif.fr.

DOI: 10.1093/nar/gkq918

PMCID: PMC3013783

PMID: 20952411 [Indexed for MEDLINE]

1799. Nucleic Acids Res. 2011 Jan;39(Database issue):D367-72. doi: 10.1093/nar/gkq906.

Epub 2010 Oct 8.

ChemProt: a disease chemical biology database.

Taboureau O(1), Nielsen SK, Audouze K, Weinhold N, Edsgärd D, Roque FS,

Kouskoumvekaki I, Bora A, Curpan R, Jensen TS, Brunak S, Oprea TI.

Author information:

(1)Department of Systems Biology, Center for Biological Sequence Analysis,

Technical University of Denmark, Lyngby, DK-2800 Denmark. otab@cbs.dtu.dk

Systems pharmacology is an emergent area that studies drug action across multiple

scales of complexity, from molecular and cellular to tissue and organism levels.

There is a critical need to develop network-based approaches to integrate the

growing body of chemical biology knowledge with network biology. Here, we report

ChemProt, a disease chemical biology database, which is based on a compilation of

multiple chemical-protein annotation resources, as well as disease-associated

protein-protein interactions (PPIs). We assembled more than 700,000 unique

chemicals with biological annotation for 30,578 proteins. We gathered over

2-million chemical-protein interactions, which were integrated in a quality

scored human PPI network of 428,429 interactions. The PPI network layer allows

for studying disease and tissue specificity through each protein complex.

ChemProt can assist in the in silico evaluation of environmental chemicals,

natural products and approved drugs, as well as the selection of new compounds

based on their activity profile against most known biological targets, including

those related to adverse drug events. Results from the disease chemical biology

database associate citalopram, an antidepressant, with osteogenesis imperfect and

leukemia and bisphenol A, an endocrine disruptor, with certain types of cancer,

respectively. The server can be accessed at

http://www.cbs.dtu.dk/services/ChemProt/.

DOI: 10.1093/nar/gkq906

PMCID: PMC3013776

PMID: 20935044 [Indexed for MEDLINE]

1800. PLoS One. 2011;6(12):e28388. doi: 10.1371/journal.pone.0028388. Epub 2011 Dec 12.

Calculating orthologs in bacteria and Archaea: a divide and conquer approach.

Halachev MR(1), Loman NJ, Pallen MJ.

Author information:

(1)School of Biosciences, University of Birmingham, Birmingham, United Kingdom.

Among proteins, orthologs are defined as those that are derived by vertical

descent from a single progenitor in the last common ancestor of their host

organisms. Our goal is to compute a complete set of protein orthologs derived

from all currently available complete bacterial and archaeal genomes. Traditional

approaches typically rely on all-against-all BLAST searching which is

prohibitively expensive in terms of hardware requirements or computational time

(requiring an estimated 18 months or more on a typical server). Here, we present

xBASE-Orth, a system for ongoing ortholog annotation, which applies a "divide and

conquer" approach and adopts a pragmatic scheme that trades accuracy for speed.

Starting at species level, xBASE-Orth carefully constructs and uses pan-genomes

as proxies for the full collections of coding sequences at each level as it

progressively climbs the taxonomic tree using the previously computed data. This

leads to a significant decrease in the number of alignments that need to be

performed, which translates into faster computation, making ortholog computation

possible on a global scale. Using xBASE-Orth, we analyzed an NCBI collection of

1,288 bacterial and 94 archaeal complete genomes with more than 4 million coding

sequences in 5 weeks and predicted more than 700 million ortholog pairs,

clustered in 175,531 orthologous groups. We have also identified sets of highly

conserved bacterial and archaeal orthologs and in so doing have highlighted

anomalies in genome annotation and in the proposed composition of the minimal

bacterial genome. In summary, our approach allows for scalable and efficient

computation of the bacterial and archaeal ortholog annotations. In addition, due

to its hierarchical nature, it is suitable for incorporating novel complete

genomes and alternative genome annotations. The computed ortholog data and a

continuously evolving set of applications based on it are integrated in the xBASE

database, available at http://www.xbase.ac.uk/.

DOI: 10.1371/journal.pone.0028388

PMCID: PMC3236195

PMID: 22174796 [Indexed for MEDLINE]

1801. PLoS One. 2011;6(11):e27836. doi: 10.1371/journal.pone.0027836. Epub 2011 Nov 16.

SProtP: a web server to recognize those short-lived proteins based on

sequence-derived features in human cells.

Song X(1), Zhou T, Jia H, Guo X, Zhang X, Han P, Sha J.

Author information:

(1)Department of Biomedical Engineering, Nanjing University of Aeronautics and

Astronautics, Nanjing, China.

Protein turnover metabolism plays important roles in cell cycle progression,

signal transduction, and differentiation. Those proteins with short half-lives

are involved in various regulatory processes. To better understand the regulation

of cell process, it is important to study the key sequence-derived factors

affecting short-lived protein degradation. Until now, most of protein half-lives

are still unknown due to the difficulties of traditional experimental methods in

measuring protein half-lives in human cells. To investigate the molecular

determinants that affect short-lived proteins, a computational method was

proposed in this work to recognize short-lived proteins based on sequence-derived

features in human cells. In this study, we have systematically analyzed many

features that perhaps correlated with short-lived protein degradation. It is

found that a large fraction of proteins with signal peptides and transmembrane

regions in human cells are of short half-lives. We have constructed an SVM-based

classifier to recognize short-lived proteins, due to the fact that short-lived

proteins play pivotal roles in the control of various cellular processes. By

employing the SVM model on human dataset, we achieved 80.8% average sensitivity

and 79.8% average specificity, respectively, on ten testing dataset (TE1-TE10).

We also obtained 89.9%, 99% and 83.9% of average accuracy on an independent

validation datasets iTE1, iTE2 and iTE3 respectively. The approach proposed in

this paper provides a valuable alternative for recognizing the short-lived

proteins in human cells, and is more accurate than the traditional N-end rule.

Furthermore, the web server SProtP (http://reprod.njmu.edu.cn/sprotp) has been

developed and is freely available for users.

DOI: 10.1371/journal.pone.0027836

PMCID: PMC3218052

PMID: 22114707 [Indexed for MEDLINE]

1802. PLoS One. 2011;6(10):e26767. doi: 10.1371/journal.pone.0026767. Epub 2011 Oct 28.

Predicting residue-residue contacts and helix-helix interactions in transmembrane

proteins using an integrative feature-based random forest approach.

Wang XF(1), Chen Z, Wang C, Yan RX, Zhang Z, Song J.

Author information:

(1)State Key Laboratory of Agrobiotechnology, College of Biological Sciences,

China Agricultural University, Beijing, China.

Integral membrane proteins constitute 25-30% of genomes and play crucial roles in

many biological processes. However, less than 1% of membrane protein structures

are in the Protein Data Bank. In this context, it is important to develop

reliable computational methods for predicting the structures of membrane

proteins. Here, we present the first application of random forest (RF) for

residue-residue contact prediction in transmembrane proteins, which we term as

TMhhcp. Rigorous cross-validation tests indicate that the built RF models provide

a more favorable prediction performance compared with two state-of-the-art

methods, i.e., TMHcon and MEMPACK. Using a strict leave-one-protein-out

jackknifing procedure, they were capable of reaching the top L/5 prediction

accuracies of 49.5% and 48.8% for two different residue contact definitions,

respectively. The predicted residue contacts were further employed to predict

interacting helical pairs and achieved the Matthew's correlation coefficients of

0.430 and 0.424, according to two different residue contact definitions,

respectively. To facilitate the academic community, the TMhhcp server has been

made freely accessible at http://protein.cau.edu.cn/tmhhcp.

DOI: 10.1371/journal.pone.0026767

PMCID: PMC3203928

PMID: 22046350 [Indexed for MEDLINE]

1803. PLoS One. 2011;6(10):e25815. doi: 10.1371/journal.pone.0025815. Epub 2011 Oct 4.

Predicting P-glycoprotein-mediated drug transport based on support vector machine

and three-dimensional crystal structure of P-glycoprotein.

Bikadi Z(1), Hazai I, Malik D, Jemnitz K, Veres Z, Hari P, Ni Z, Loo TW, Clarke

DM, Hazai E, Mao Q.

Author information:

(1)Virtua Drug Ltd, Budapest, Hungary.

Human P-glycoprotein (P-gp) is an ATP-binding cassette multidrug transporter that

confers resistance to a wide range of chemotherapeutic agents in cancer cells by

active efflux of the drugs from cells. P-gp also plays a key role in limiting

oral absorption and brain penetration and in facilitating biliary and renal

elimination of structurally diverse drugs. Thus, identification of drugs or new

molecular entities to be P-gp substrates is of vital importance for predicting

the pharmacokinetics, efficacy, safety, or tissue levels of drugs or drug

candidates. At present, publicly available, reliable in silico models predicting

P-gp substrates are scarce. In this study, a support vector machine (SVM) method

was developed to predict P-gp substrates and P-gp-substrate interactions, based

on a training data set of 197 known P-gp substrates and non-substrates collected

from the literature. We showed that the SVM method had a prediction accuracy of

approximately 80% on an independent external validation data set of 32 compounds.

A homology model of human P-gp based on the X-ray structure of mouse P-gp as a

template has been constructed. We showed that molecular docking to the P-gp

structures successfully predicted the geometry of P-gp-ligand complexes. Our SVM

prediction and the molecular docking methods have been integrated into a free web

server (http://pgp.althotas.com), which allows the users to predict whether a

given compound is a P-gp substrate and how it binds to and interacts with P-gp.

Utilization of such a web server may prove valuable for both rational drug design

and screening.

DOI: 10.1371/journal.pone.0025815

PMCID: PMC3186768

PMID: 21991360 [Indexed for MEDLINE]

1804. PLoS One. 2011;6(10):e25560. doi: 10.1371/journal.pone.0025560. Epub 2011 Oct 3.

Computational prediction of heme-binding residues by exploiting residue

interaction network.

Liu R(1), Hu J.

Author information:

(1)Department of Computer Science and Engineering, University of South Carolina,

Columbia, South Carolina, United States of America.

Computational identification of heme-binding residues is beneficial for

predicting and designing novel heme proteins. Here we proposed a novel method for

heme-binding residue prediction by exploiting topological properties of these

residues in the residue interaction networks derived from three-dimensional

structures. Comprehensive analysis showed that key residues located in

heme-binding regions are generally associated with the nodes with higher degree,

closeness and betweenness, but lower clustering coefficient in the network.

HemeNet, a support vector machine (SVM) based predictor, was developed to

identify heme-binding residues by combining topological features with existing

sequence and structural features. The results showed that incorporation of

network-based features significantly improved the prediction performance. We also

compared the residue interaction networks of heme proteins before and after heme

binding and found that the topological features can well characterize the

heme-binding sites of apo structures as well as those of holo structures, which

led to reliable performance improvement as we applied HemeNet to predicting the

binding residues of proteins in the heme-free state. HemeNet web server is freely

accessible at http://mleg.cse.sc.edu/hemeNet/.

DOI: 10.1371/journal.pone.0025560

PMCID: PMC3184988

PMID: 21991319 [Indexed for MEDLINE]

1805. PLoS One. 2011;6(9):e24914. doi: 10.1371/journal.pone.0024914. Epub 2011 Sep 22.

Persistence and availability of Web services in computational biology.

Schultheiss SJ(1), Münch MC, Andreeva GD, Rätsch G.

Author information:

(1)Machine Learning in Biology Research Group, Friedrich Miescher Laboratory of

the Max Planck Society, Tübingen, Germany. sebastian.schultheiss@tue.mpg.de

We have conducted a study on the long-term availability of bioinformatics Web

services: an observation of 927 Web services published in the annual Nucleic

Acids Research Web Server Issues between 2003 and 2009. We found that 72% of Web

sites are still available at the published addresses, only 9% of services are

completely unavailable. Older addresses often redirect to new pages. We checked

the functionality of all available services: for 33%, we could not test

functionality because there was no example data or a related problem; 13% were

truly no longer working as expected; we could positively confirm functionality

only for 45% of all services. Additionally, we conducted a survey among 872 Web

Server Issue corresponding authors; 274 replied. 78% of all respondents indicate

their services have been developed solely by students and researchers without a

permanent position. Consequently, these services are in danger of falling into

disrepair after the original developers move to another institution, and indeed,

for 24% of services, there is no plan for maintenance, according to the

respondents. We introduce a Web service quality scoring system that correlates

with the number of citations: services with a high score are cited 1.8 times more

often than low-scoring services. We have identified key characteristics that are

predictive of a service's survival, providing reviewers, editors, and Web service

developers with the means to assess or improve Web services. A Web service

conforming to these criteria receives more citations and provides more reliable

service for its users. The most effective way of ensuring continued access to a

service is a persistent Web address, offered either by the publishing journal, or

created on the authors' own initiative, for example at http://bioweb.me. The

community would benefit the most from a policy requiring any source code needed

to reproduce results to be deposited in a public repository.

DOI: 10.1371/journal.pone.0024914

PMCID: PMC3178567

PMID: 21966383 [Indexed for MEDLINE]

1806. PLoS One. 2011;6(9):e24583. doi: 10.1371/journal.pone.0024583. Epub 2011 Sep 15.

MultiMiTar: a novel multi objective optimization based miRNA-target prediction

method.

Mitra R(1), Bandyopadhyay S.

Author information:

(1)Machine Intelligence Unit, Indian Statistical Institute, Kolkata, West Bengal,

India.

BACKGROUND: Machine learning based miRNA-target prediction algorithms often fail

to obtain a balanced prediction accuracy in terms of both sensitivity and

specificity due to lack of the gold standard of negative examples,

miRNA-targeting site context specific relevant features and efficient feature

selection process. Moreover, all the sequence, structure and machine learning

based algorithms are unable to distribute the true positive predictions

preferentially at the top of the ranked list; hence the algorithms become

unreliable to the biologists. In addition, these algorithms fail to obtain

considerable combination of precision and recall for the target transcripts that

are translationally repressed at protein level.

METHODOLOGY/PRINCIPAL FINDING: In the proposed article, we introduce an efficient

miRNA-target prediction system MultiMiTar, a Support Vector Machine (SVM) based

classifier integrated with a multiobjective metaheuristic based feature selection

technique. The robust performance of the proposed method is mainly the result of

using high quality negative examples and selection of biologically relevant

miRNA-targeting site context specific features. The features are selected by

using a novel feature selection technique AMOSA-SVM, that integrates the multi

objective optimization technique Archived Multi-Objective Simulated Annealing

(AMOSA) and SVM.

CONCLUSIONS/SIGNIFICANCE: MultiMiTar is found to achieve much higher Matthew's

correlation coefficient (MCC) of 0.583 and average class-wise accuracy (ACA) of

0.8 compared to the others target prediction methods for a completely independent

test data set. The obtained MCC and ACA values of these algorithms range from

-0.269 to 0.155 and 0.321 to 0.582, respectively. Moreover, it shows a more

balanced result in terms of precision and sensitivity (recall) for the

translationally repressed data set as compared to all the other existing methods.

An important aspect is that the true positive predictions are distributed

preferentially at the top of the ranked list that makes MultiMiTar reliable for

the biologists. MultiMiTar is now available as an online tool at

www.isical.ac.in/~bioinfo\_miu/multimitar.htm. MultiMiTar software can be

downloaded from www.isical.ac.in/~bioinfo\_miu/multimitar-download.htm.

DOI: 10.1371/journal.pone.0024583

PMCID: PMC3174180

PMID: 21949731 [Indexed for MEDLINE]

1807. PLoS One. 2011;6(9):e24756. doi: 10.1371/journal.pone.0024756. Epub 2011 Sep 15.

iDNA-Prot: identification of DNA binding proteins using random forest with grey

model.

Lin WZ(1), Fang JA, Xiao X, Chou KC.

Author information:

(1)Information Science and Technology School, Donghua University, Shanghai,

China.

DNA-binding proteins play crucial roles in various cellular processes. Developing

high throughput tools for rapidly and effectively identifying DNA-binding

proteins is one of the major challenges in the field of genome annotation.

Although many efforts have been made in this regard, further effort is needed to

enhance the prediction power. By incorporating the features into the general form

of pseudo amino acid composition that were extracted from protein sequences via

the "grey model" and by adopting the random forest operation engine, we proposed

a new predictor, called iDNA-Prot, for identifying uncharacterized proteins as

DNA-binding proteins or non-DNA binding proteins based on their amino acid

sequences information alone. The overall success rate by iDNA-Prot was 83.96%

that was obtained via jackknife tests on a newly constructed stringent benchmark

dataset in which none of the proteins included has ≥25% pairwise sequence

identity to any other in a same subset. In addition to achieving high success

rate, the computational time for iDNA-Prot is remarkably shorter in comparison

with the relevant existing predictors. Hence it is anticipated that iDNA-Prot may

become a useful high throughput tool for large-scale analysis of DNA-binding

proteins. As a user-friendly web-server, iDNA-Prot is freely accessible to the

public at the web-site on http://icpr.jci.edu.cn/bioinfo/iDNA-Prot or

http://www.jci-bioinfo.cn/iDNA-Prot. Moreover, for the convenience of the vast

majority of experimental scientists, a step-by-step guide is provided on how to

use the web-server to get the desired results.

DOI: 10.1371/journal.pone.0024756

PMCID: PMC3174210

PMID: 21935457 [Indexed for MEDLINE]

1808. PLoS One. 2011;6(9):e24039. doi: 10.1371/journal.pone.0024039. Epub 2011 Sep 13.

Identification of mannose interacting residues using local composition.

Agarwal S(1), Mishra NK, Singh H, Raghava GP.

Author information:

(1)Institute of Microbial Technology, Chandigarh, India.

BACKGROUND: Mannose binding proteins (MBPs) play a vital role in several

biological functions such as defense mechanisms. These proteins bind to mannose

on the surface of a wide range of pathogens and help in eliminating these

pathogens from our body. Thus, it is important to identify mannose interacting

residues (MIRs) in order to understand mechanism of recognition of pathogens by

MBPs.

RESULTS: This paper describes modules developed for predicting MIRs in a protein.

Support vector machine (SVM) based models have been developed on 120 mannose

binding protein chains, where no two chains have more than 25% sequence

similarity. SVM models were developed on two types of datasets: 1) main dataset

consists of 1029 mannose interacting and 1029 non-interacting residues, 2)

realistic dataset consists of 1029 mannose interacting and 10320 non-interacting

residues. In this study, firstly, we developed standard modules using binary and

PSSM profile of patterns and got maximum MCC around 0.32. Secondly, we developed

SVM modules using composition profile of patterns and achieved maximum MCC around

0.74 with accuracy 86.64% on main dataset. Thirdly, we developed a model on a

realistic dataset and achieved maximum MCC of 0.62 with accuracy 93.08%. Based on

this study, a standalone program and web server have been developed for

predicting mannose interacting residues in proteins

(http://www.imtech.res.in/raghava/premier/).

CONCLUSIONS: Compositional analysis of mannose interacting and non-interacting

residues shows that certain types of residues are preferred in mannose

interaction. It was also observed that residues around mannose interacting

residues have a preference for certain types of residues. Composition of

patterns/peptide/segment has been used for predicting MIRs and achieved

reasonable high accuracy. It is possible that this novel strategy may be

effective to predict other types of interacting residues. This study will be

useful in annotating the function of protein as well as in understanding the role

of mannose in the immune system.

DOI: 10.1371/journal.pone.0024039

PMCID: PMC3172211

PMID: 21931639 [Indexed for MEDLINE]

1809. PLoS One. 2011;6(8):e23505. doi: 10.1371/journal.pone.0023505. Epub 2011 Aug 15.

NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based

on sequence-derived features.

Wang P(1), Xiao X, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China.

Nuclear receptors (NRs) are one of the most abundant classes of transcriptional

regulators in animals. They regulate diverse functions, such as homeostasis,

reproduction, development and metabolism. Therefore, NRs are a very important

target for drug development. Nuclear receptors form a superfamily of

phylogenetically related proteins and have been subdivided into different

subfamilies due to their domain diversity. In this study, a two-level predictor,

called NR-2L, was developed that can be used to identify a query protein as a

nuclear receptor or not based on its sequence information alone; if it is, the

prediction will be automatically continued to further identify it among the

following seven subfamilies: (1) thyroid hormone like (NR1), (2) HNF4-like (NR2),

(3) estrogen like, (4) nerve growth factor IB-like (NR4), (5) fushi tarazu-F1

like (NR5), (6) germ cell nuclear factor like (NR6), and (7) knirps like (NR0).

The identification was made by the Fuzzy K nearest neighbor (FK-NN) classifier

based on the pseudo amino acid composition formed by incorporating various

physicochemical and statistical features derived from the protein sequences, such

as amino acid composition, dipeptide composition, complexity factor, and

low-frequency Fourier spectrum components. As a demonstration, it was shown

through some benchmark datasets derived from the NucleaRDB and UniProt with low

redundancy that the overall success rates achieved by the jackknife test were

about 93% and 89% in the first and second level, respectively. The high success

rates indicate that the novel two-level predictor can be a useful vehicle for

identifying NRs and their subfamilies. As a user-friendly web server, NR-2L is

freely accessible at either http://icpr.jci.edu.cn/bioinfo/NR2L or

http://www.jci-bioinfo.cn/NR2L. Each job submitted to NR-2L can contain up to 500

query protein sequences and be finished in less than 2 minutes. The less the

number of query proteins is, the shorter the time will usually be. All the

program codes for NR-2L are available for non-commercial purpose upon request.

DOI: 10.1371/journal.pone.0023505

PMCID: PMC3156231

PMID: 21858146 [Indexed for MEDLINE]

1810. PLoS One. 2011;6(7):e22270. doi: 10.1371/journal.pone.0022270. Epub 2011 Jul 20.

Exploiting publicly available biological and biochemical information for the

discovery of novel short linear motifs.

Sayadi A(1), Briganti L, Tramontano A, Via A.

Author information:

(1)Department of Physics, Sapienza University of Rome, Rome, Italy.

The function of proteins is often mediated by short linear segments of their

amino acid sequence, called Short Linear Motifs or SLiMs, the identification of

which can provide important information about a protein function. However, the

short length of the motifs and their variable degree of conservation makes their

identification hard since it is difficult to correctly estimate the statistical

significance of their occurrence. Consequently, only a small fraction of them

have been discovered so far. We describe here an approach for the discovery of

SLiMs based on their occurrence in evolutionarily unrelated proteins belonging to

the same biological, signalling or metabolic pathway and give specific examples

of its effectiveness in both rediscovering known motifs and in discovering novel

ones. An automatic implementation of the procedure, available for download,

allows significant motifs to be identified, automatically annotated with

functional, evolutionary and structural information and organized in a database

that can be inspected and queried. An instance of the database populated with

pre-computed data on seven organisms is accessible through a publicly available

server and we believe it constitutes by itself a useful resource for the life

sciences (http://www.biocomputing.it/modipath).

DOI: 10.1371/journal.pone.0022270

PMCID: PMC3140502

PMID: 21799808 [Indexed for MEDLINE]

1811. PLoS One. 2011;6(7):e21800. doi: 10.1371/journal.pone.0021800. Epub 2011 Jul 18.

REVIGO summarizes and visualizes long lists of gene ontology terms.

Supek F(1), Bošnjak M, Škunca N, Šmuc T.

Author information:

(1)Division of Electronics, Rudjer Boskovic Institute, Zagreb, Croatia.

fran.supek@irb.hr

Outcomes of high-throughput biological experiments are typically interpreted by

statistical testing for enriched gene functional categories defined by the Gene

Ontology (GO). The resulting lists of GO terms may be large and highly redundant,

and thus difficult to interpret.REVIGO is a Web server that summarizes long,

unintelligible lists of GO terms by finding a representative subset of the terms

using a simple clustering algorithm that relies on semantic similarity measures.

Furthermore, REVIGO visualizes this non-redundant GO term set in multiple ways to

assist in interpretation: multidimensional scaling and graph-based visualizations

accurately render the subdivisions and the semantic relationships in the data,

while treemaps and tag clouds are also offered as alternative views. REVIGO is

freely available at http://revigo.irb.hr/.

DOI: 10.1371/journal.pone.0021800

PMCID: PMC3138752

PMID: 21789182 [Indexed for MEDLINE]

1812. PLoS One. 2011;6(7):e21270. doi: 10.1371/journal.pone.0021270. Epub 2011 Jul 5.

GenExp: an interactive web-based genomic DAS client with client-side data

rendering.

Gel Moreno B(1), Messeguer Peypoch X.

Author information:

(1)Software Department, UPC-BarcelonaTech, Barcelona, Spain. bgel@lsi.upc.edu

BACKGROUND: The Distributed Annotation System (DAS) offers a standard protocol

for sharing and integrating annotations on biological sequences. There are more

than 1000 DAS sources available and the number is steadily increasing. Clients

are an essential part of the DAS system and integrate data from several

independent sources in order to create a useful representation to the user. While

web-based DAS clients exist, most of them do not have direct interaction

capabilities such as dragging and zooming with the mouse.

RESULTS: Here we present GenExp, a web based and fully interactive visual DAS

client. GenExp is a genome oriented DAS client capable of creating informative

representations of genomic data zooming out from base level to complete

chromosomes. It proposes a novel approach to genomic data rendering and uses the

latest HTML5 web technologies to create the data representation inside the client

browser. Thanks to client-side rendering most position changes do not need a

network request to the server and so responses to zooming and panning are almost

immediate. In GenExp it is possible to explore the genome intuitively moving it

with the mouse just like geographical map applications. Additionally, in GenExp

it is possible to have more than one data viewer at the same time and to save the

current state of the application to revisit it later on.

CONCLUSIONS: GenExp is a new interactive web-based client for DAS and addresses

some of the short-comings of the existing clients. It uses client-side data

rendering techniques resulting in easier genome browsing and exploration. GenExp

is open source under the GPL license and it is freely available at

http://gralggen.lsi.upc.edu/recerca/genexp.

DOI: 10.1371/journal.pone.0021270

PMCID: PMC3130039

PMID: 21750706 [Indexed for MEDLINE]

1813. PLoS One. 2011;6(6):e20592. doi: 10.1371/journal.pone.0020592. Epub 2011 Jun 17.

A multi-label classifier for predicting the subcellular localization of

gram-negative bacterial proteins with both single and multiple sites.

Xiao X(1), Wu ZC, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China.

xiaoxuan0326@yahoo.com.cn

Prediction of protein subcellular localization is a challenging problem,

particularly when the system concerned contains both singleplex and multiplex

proteins. In this paper, by introducing the "multi-label scale" and hybridizing

the information of gene ontology with the sequential evolution information, a

novel predictor called iLoc-Gneg is developed for predicting the subcellular

localization of gram-positive bacterial proteins with both single-location and

multiple-location sites. For facilitating comparison, the same stringent

benchmark dataset used to estimate the accuracy of Gneg-mPLoc was adopted to

demonstrate the power of iLoc-Gneg. The dataset contains 1,392 gram-negative

bacterial proteins classified into the following eight locations: (1) cytoplasm,

(2) extracellular, (3) fimbrium, (4) flagellum, (5) inner membrane, (6) nucleoid,

(7) outer membrane, and (8) periplasm. Of the 1,392 proteins, 1,328 are each with

only one subcellular location and the other 64 are each with two subcellular

locations, but none of the proteins included has pairwise sequence identity to

any other in a same subset (subcellular location). It was observed that the

overall success rate by jackknife test on such a stringent benchmark dataset by

iLoc-Gneg was over 91%, which is about 6% higher than that by Gneg-mPLoc. As a

user-friendly web-server, iLoc-Gneg is freely accessible to the public at

http://icpr.jci.edu.cn/bioinfo/iLoc-Gneg. Meanwhile, a step-by-step guide is

provided on how to use the web-server to get the desired results. Furthermore,

for the user's convenience, the iLoc-Gneg web-server also has the function to

accept the batch job submission, which is not available in the existing version

of Gneg-mPLoc web-server. It is anticipated that iLoc-Gneg may become a useful

high throughput tool for Molecular Cell Biology, Proteomics, System Biology, and

Drug Development.

DOI: 10.1371/journal.pone.0020592

PMCID: PMC3117797

PMID: 21698097 [Indexed for MEDLINE]

1814. PLoS One. 2011;6(5):e20025. doi: 10.1371/journal.pone.0020025. Epub 2011 May 25.

Discovery of protein phosphorylation motifs through exploratory data analysis.

Chen YC(1), Aguan K, Yang CW, Wang YT, Pal NR, Chung IF.

Author information:

(1)Institute of Biomedical Informatics, National Yang-Ming University, Taipei,

Taiwan.

BACKGROUND: The need for efficient algorithms to uncover biologically relevant

phosphorylation motifs has become very important with rapid expansion of the

proteomic sequence database along with a plethora of new information on

phosphorylation sites. Here we present a novel unsupervised method, called Motif

Finder (in short, F-Motif) for identification of phosphorylation motifs. F-Motif

uses clustering of sequence information represented by numerical features that

exploit the statistical information hidden in some foreground data. Furthermore,

these identified motifs are then filtered to find "actual" motifs with

statistically significant motif scores.

RESULTS AND DISCUSSION: We have applied F-Motif to several new and existing data

sets and compared its performance with two well known state-of-the-art methods.

In almost all cases F-Motif could identify all statistically significant motifs

extracted by the state-of-the-art methods. More importantly, in addition to this,

F-Motif uncovers several novel motifs. We have demonstrated using clues from the

literature that most of these new motifs discovered by F-Motif are indeed novel.

We have also found some interesting phenomena. For example, for CK2 kinase, the

conserved sites appear only on the right side of S. However, for CDK kinase, the

adjacent site on the right of S is conserved with residue P. In addition, three

different encoding methods, including a novel position contrast matrix (PCM) and

the simplest binary coding, are used and the ability of F-motif to discover

motifs remains quite robust with respect to encoding schemes.

CONCLUSIONS: An iterative algorithm proposed here uses exploratory data analysis

to discover motifs from phosphorylated data. The effectiveness of F-Motif has

been demonstrated using several real data sets as well as using a synthetic data

set. The method is quite general in nature and can be used to find other types of

motifs also. We have also provided a server for F-Motif at

http://f-motif.classcloud.org/, http://bio.classcloud.org/f-motif/ or

http://ymu.classcloud.org/f-motif/.

DOI: 10.1371/journal.pone.0020025

PMCID: PMC3102080

PMID: 21647451 [Indexed for MEDLINE]

1815. Protein Pept Lett. 2011 Jan;18(1):58-63.

Identify Golgi protein types with modified Mahalanobis discriminant algorithm and

pseudo amino acid composition.

Ding H(1), Liu L, Guo FB, Huang J, Lin H.

Author information:

(1)Key Laboratory for NeuroInformation of Ministry of Education, School of Life

Science and Technology, University of Electronic Science and Technology of China,

Chengdu, 610054, China.

The Golgi apparatus is an important eukaryotic organelle. Successful prediction

of Golgi protein types can provide valuable information for elucidating protein

functions involved in various biological processes. In this work, a method is

proposed by combining a special mode of pseudo amino acid composition (increment

of diversity) with the modified Mahalanobis discriminant for predicting Golgi

protein types. The benchmark dataset used to train the predictor thus formed

contains 95 Golgi proteins in which none of proteins included has ≥40% pairwise

sequence identity to any other. The accuracy obtained by the jackknife test was

74.7%, with the ROC curve of 0.772 in identifying cis-Golgi proteins and

trans-Golgi proteins. Subsequently, the method was extended to discriminate

cis-Golgi network proteins from cis-Golgi network membrane proteins and

trans-Golgi network proteins from trans-Golgi network membrane proteins,

respectively. The accuracies thus obtained were 76.1% and 83.7%, respectively.

These results indicate that our method may become a useful tool in the relevant

areas. As a user-friendly web-server, the predictor is freely accessible at

http://immunet.cn/SubGolgi/.

PMID: 20955168 [Indexed for MEDLINE]

1816. Proteins. 2011;79 Suppl 10:59-73. doi: 10.1002/prot.23181. Epub 2011 Oct 14.

CASP9 assessment of free modeling target predictions.

Kinch L(1), Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV.

Author information:

(1)Howard Hughes Medical Institute, University of Texas, Southwestern Medical

Center, Dallas, TX 75390-9050, USA. . lkinch@chop.swmed.edu

We present an overview of the ninth round of Critical Assessment of Protein

Structure Prediction (CASP9) "Template free modeling" category (FM). Prediction

models were evaluated using a combination of established structural and sequence

comparison measures and a novel automated method designed to mimic manual

inspection by capturing both global and local structural features. These scores

were compared to those assigned manually over a diverse subset of target domains.

Scores were combined to compare overall performance of participating groups and

to estimate rank significance. Moreover, we discuss a few examples of free

modeling targets to highlight the progress and bottlenecks of current prediction

methods. Notably, a server prediction model for a single target (T0581) improved

significantly over the closest structure template (44% GDT increase). This

accomplishment represents the "winner" of the CASP9 FM category. A number of

human expert groups submitted slight variations of this model, highlighting a

trend for human experts to act as "meta predictors" by correctly selecting among

models produced by the top-performing automated servers. The details of

evaluation are available at http://prodata.swmed.edu/CASP9/ .

Copyright © 2011 Wiley-Liss, Inc.

DOI: 10.1002/prot.23181

PMCID: PMC3226891

PMID: 21997521 [Indexed for MEDLINE]

1817. Proteins. 2011;79 Suppl 10:161-71. doi: 10.1002/prot.23175. Epub 2011 Oct 11.

RaptorX: exploiting structure information for protein alignment by statistical

inference.

Peng J(1), Xu J.

Author information:

(1)Toyota Technological Institute at Chicago, 6045 S. Kenwood Avenue, Chicago, IL

60637, USA.

This work presents RaptorX, a statistical method for template-based protein

modeling that improves alignment accuracy by exploiting structural information in

a single or multiple templates. RaptorX consists of three major components:

single-template threading, alignment quality prediction, and multiple-template

threading. This work summarizes the methods used by RaptorX and presents its

CASP9 result analysis, aiming to identify major bottlenecks with RaptorX and

template-based modeling and hopefully directions for further study. Our results

show that template structural information helps a lot with both single-template

and multiple-template protein threading especially when closely-related templates

are unavailable, and there is still large room for improvement in both alignment

and template selection. The RaptorX web server is available at

http://raptorx.uchicago.edu.

Copyright © 2011 Wiley-Liss, Inc.

DOI: 10.1002/prot.23175

PMCID: PMC3226909

PMID: 21987485 [Indexed for MEDLINE]

1818. RNA. 2011 Jan;17(1):27-38. doi: 10.1261/rna.2394511. Epub 2010 Nov 22.

Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins.

Sperschneider J(1), Datta A, Wise MJ.

Author information:

(1)School of Computer Science and Software Engineering, University of Western

Australia, Perth WA 6009, Australia. janaspe@csse.uwa.edu.au

Pseudoknots are an essential feature of RNA tertiary structures. Simple H-type

pseudoknots have been studied extensively in terms of biological functions,

computational prediction, and energy models. Intramolecular kissing hairpins are

a more complex and biologically important type of pseudoknot in which two hairpin

loops form base pairs. They are hard to predict using free energy minimization

due to high computational requirements. Heuristic methods that allow arbitrary

pseudoknots strongly depend on the quality of energy parameters, which are not

yet available for complex pseudoknots. We present an extension of the heuristic

pseudoknot prediction algorithm DotKnot, which covers H-type pseudoknots and

intramolecular kissing hairpins. Our framework allows for easy integration of

advanced H-type pseudoknot energy models. For a test set of RNA sequences

containing kissing hairpins and other types of pseudoknot structures, DotKnot

outperforms competing methods from the literature. DotKnot is available as a web

server under http://dotknot.csse.uwa.edu.au.

DOI: 10.1261/rna.2394511

PMCID: PMC3004063

PMID: 21098139 [Indexed for MEDLINE]

1819. RNA Biol. 2011 Jan-Feb;8(1):11-3. Epub 2011 Jan 1.

ARNold: a web tool for the prediction of Rho-independent transcription

terminators.

Naville M(1), Ghuillot-Gaudeffroy A, Marchais A, Gautheret D.

Author information:

(1)Univ. Paris-Sud 11, CNRS, UMR8621, Institut de Génétique et Microbiologie,

Orsay Cedex, France.

Rho-independent termination is a major mechanism of transcriptional arrest in

bacteria that controls both normal 3' termination and a wide array of regulatory

attenuation events. Detecting Rho-independent terminators is an obliged step in

the annotation of bacterial operons. Yet, while several efficient algorithms are

available for this purpose, there is no freely available web site enabling a

rapid scanning of raw genomic sequence for the presence of terminators. Here we

implemented such a web server, which combines two published prediction

algorithms, Erpin and RNAmotif, and performs nearly as well as more complex

procedures while being accessible to the non specialist. The ARNold Web server is

available at : http://rna.igmors.u-psud.fr/toolbox/arnold/

PMID: 21282983 [Indexed for MEDLINE]

1820. BMC Bioinformatics. 2010 Dec 17;11:600. doi: 10.1186/1471-2105-11-600.

SNPexp - A web tool for calculating and visualizing correlation between HapMap

genotypes and gene expression levels.

Holm K(1), Melum E, Franke A, Karlsen TH.

Author information:

(1)Norwegian PSC Research Center, Clinic for Specialized Medicine and Surgery,

Oslo University Hospital Rikshospitalet, Norway.

BACKGROUND: Expression levels for 47294 transcripts in lymphoblastoid cell lines

from all 270 HapMap phase II individuals, and genotypes (both HapMap phase II and

III) of 3.96 million single nucleotide polymorphisms (SNPs) in the same

individuals are publicly available. We aimed to generate a user-friendly web

based tool for visualization of the correlation between SNP genotypes within a

specified genomic region and a gene of interest, which is also well-known as an

expression quantitative trait locus (eQTL) analysis.

RESULTS: SNPexp is implemented as a server-side script, and publicly available on

this website: http://tinyurl.com/snpexp. Correlation between genotype and

transcript expression levels are calculated by performing linear regression and

the Wald test as implemented in PLINK and visualized using the UCSC Genome

Browser. Validation of SNPexp using previously published eQTLs yielded comparable

results.

CONCLUSIONS: SNPexp provides a convenient and platform-independent way to

calculate and visualize the correlation between HapMap genotypes within a

specified genetic region anywhere in the genome and gene expression levels. This

allows for investigation of both cis and trans effects. The web interface and

utilization of publicly available and widely used software resources makes it an

attractive supplement to more advanced bioinformatic tools. For the advanced user

the program can be used on a local computer on custom datasets.

DOI: 10.1186/1471-2105-11-600

PMCID: PMC3022629

PMID: 21167019 [Indexed for MEDLINE]

1821. J Med Internet Res. 2010 Dec 17;12(4):e71. doi: 10.2196/jmir.1338.

An online community improves adherence in an internet-mediated walking program.

Part 1: results of a randomized controlled trial.

Richardson CR(1), Buis LR, Janney AW, Goodrich DE, Sen A, Hess ML, Mehari KS,

Fortlage LA, Resnick PJ, Zikmund-Fisher BJ, Strecher VJ, Piette JD.

Author information:

(1)Department of Family Medicine, University of Michigan, Ann Arbor, MI

48104-1213, United States. caroli@umich.edu

BACKGROUND: Approximately half of American adults do not meet recommended

physical activity guidelines. Face-to-face lifestyle interventions improve health

outcomes but are unlikely to yield population-level improvements because they can

be difficult to disseminate, expensive to maintain, and inconvenient for the

recipient. In contrast, Internet-based behavior change interventions can be

disseminated widely at a lower cost. However, the impact of some

Internet-mediated programs is limited by high attrition rates. Online communities

that allow participants to communicate with each other by posting and reading

messages may decrease participant attrition.

OBJECTIVE: Our objective was to measure the impact of adding online community

features to an Internet-mediated walking program on participant attrition and

average daily step counts.

METHODS: This randomized controlled trial included sedentary, ambulatory adults

who used email regularly and had at least 1 of the following: overweight (body

mass index [BMI] ≥ 25), type 2 diabetes, or coronary artery disease. All

participants (n = 324) wore enhanced pedometers throughout the 16-week

intervention and uploaded step-count data to the study server. Participants could

log in to the study website to view graphs of their walking progress,

individually-tailored motivational messages, and weekly calculated goals.

Participants were randomized to 1 of 2 versions of a Web-based walking program.

Those randomized to the "online community" arm could post and read messages with

other participants while those randomized to the "no online community" arm could

not read or post messages. The main outcome measures were participant attrition

and average daily step counts over 16 weeks. Multiple regression analyses

assessed the effect of the online community access controlling for age, sex,

disease status, BMI, and baseline step counts.

RESULTS: Both arms significantly increased their average daily steps between

baseline and the end of the intervention period, but there were no significant

differences in increase in step counts between arms using either

intention-to-treat or completers analysis. In the intention-to-treat analysis,

the average step count increase across both arms was 1888 ± 2400 steps. The

percentage of completers was 13% higher in the online community arm than the no

online community arm (online community arm, 79%, no online community arm, 66%, P

= .02). In addition, online community arm participants remained engaged in the

program longer than no online community arm participants (hazard ratio = 0.47,

95% CI = 0.25 - 0.90, P = .02). Participants with lower baseline social support

posted more messages to the online community (P < .001) and viewed more posts (P

< .001) than participants with higher baseline social support.

CONCLUSION: Adding online community features to an Internet-mediated walking

program did not increase average daily step counts but did reduce participant

attrition. Participants with low baseline social support used the online

community features more than those with high baseline social support. Thus,

online communities may be a promising approach to reducing attrition from online

health behavior change interventions, particularly in populations with low social

support.

TRIAL REGISTRATION: NCT00729040; http://clinicaltrials.gov/ct2/show/NCT00729040

(Archived by WebCite at http://www.webcitation.org/5v1VH3n0A).

DOI: 10.2196/jmir.1338

PMCID: PMC3056526

PMID: 21169160 [Indexed for MEDLINE]

1822. Bioinformatics. 2010 Dec 15;26(24):3127-8. doi: 10.1093/bioinformatics/btq601.

Epub 2010 Oct 28.

PPO: predictor for prokaryotic operons.

Chuang LY(1), Tsai JH, Yang CH.

Author information:

(1)Department of Chemical Engineering & Institute of Biotechnology and Chemical

Engineering, I-Shou University, Kaohsiung, Taiwan.

SUMMARY: We present an operon predictor for prokaryotic operons (PPO), which can

predict operons in the entire prokaryotic genome. The prediction algorithm used

in PPO allows the user to select binary particle swarm optimization (BPSO), a

genetic algorithm (GA) or some other methods introduced in the literature to

predict operons. The operon predictor on our web server and the provided database

are easy to access and use. The main features offered are: (i) selection of the

prediction algorithm; (ii) adjustable parameter settings of the prediction

algorithm; (iii) graphic visualization of results; (iv) integrated database

queries; (v) listing of experimentally verified operons; and (vi) related tools.

AVAILABILITY AND IMPLEMENTATION: PPO is freely available at

http://bio.kuas.edu.tw/PPO/.

DOI: 10.1093/bioinformatics/btq601

PMID: 21030461 [Indexed for MEDLINE]

1823. Bioinformatics. 2010 Dec 15;26(24):3119-24. doi: 10.1093/bioinformatics/btq599.

Epub 2010 Oct 22.

The GMOD Drupal bioinformatic server framework.

Papanicolaou A(1), Heckel DG.

Author information:

(1)Centre for Conservation and Ecology, University of Exeter in Cornwall, Penryn,

UK. alexie@butterflybase.org

MOTIVATION: Next-generation sequencing technologies have led to the widespread

use of -omic applications. As a result, there is now a pronounced bioinformatic

bottleneck. The general model organism database (GMOD) tool kit (http://gmod.org)

has produced a number of resources aimed at addressing this issue. It lacks,

however, a robust online solution that can deploy heterogeneous data and software

within a Web content management system (CMS).

RESULTS: We present a bioinformatic framework for the Drupal CMS. It consists of

three modules. First, GMOD-DBSF is an application programming interface module

for the Drupal CMS that simplifies the programming of bioinformatic Drupal

modules. Second, the Drupal Bioinformatic Software Bench (biosoftware\_bench)

allows for a rapid and secure deployment of bioinformatic software. An innovative

graphical user interface (GUI) guides both use and administration of the

software, including the secure provision of pre-publication datasets. Third, we

present genes4all\_experiment, which exemplifies how our work supports the wider

research community.

CONCLUSION: Given the infrastructure presented here, the Drupal CMS may become a

powerful new tool set for bioinformaticians. The GMOD-DBSF base module is an

expandable community resource that decreases development time of Drupal modules

for bioinformatics. The biosoftware\_bench module can already enhance biologists'

ability to mine their own data. The genes4all\_experiment module has already been

responsible for archiving of more than 150 studies of RNAi from Lepidoptera,

which were previously unpublished.

AVAILABILITY AND IMPLEMENTATION: Implemented in PHP and Perl. Freely available

under the GNU Public License 2 or later from http://gmod-dbsf.googlecode.com.

DOI: 10.1093/bioinformatics/btq599

PMCID: PMC2995126

PMID: 20971988 [Indexed for MEDLINE]

1824. PLoS One. 2010 Dec 13;5(12):e15305. doi: 10.1371/journal.pone.0015305.

TogoDoc server/client system: smart recommendation and efficient management of

life science literature.

Iwasaki W(1), Yamamoto Y, Takagi T.

Author information:

(1)Department of Computational Biology, University of Tokyo, Kashiwa, Japan.

iwasaki@k.u-tokyo.ac.jp

In this paper, we describe a server/client literature management system

specialized for the life science domain, the TogoDoc system (Togo, pronounced

Toe-Go, is a romanization of a Japanese word for integration). The server and the

client program cooperate closely over the Internet to provide life scientists

with an effective literature recommendation service and efficient literature

management. The content-based and personalized literature recommendation helps

researchers to isolate interesting papers from the "tsunami" of literature, in

which, on average, more than one biomedical paper is added to MEDLINE every

minute. Because researchers these days need to cover updates of much wider topics

to generate hypotheses using massive datasets obtained from public databases or

omics experiments, the importance of having an effective literature

recommendation service is rising. The automatic recommendation is based on the

content of personal literature libraries of electronic PDF papers. The client

program automatically analyzes these files, which are sometimes deeply buried in

storage disks of researchers' personal computers. Just saving PDF papers to the

designated folders makes the client program automatically analyze and retrieve

metadata, rename file names, synchronize the data to the server, and receive the

recommendation lists of newly published papers, thus accomplishing effortless

literature management. In addition, the tag suggestion and associative search

functions are provided for easy classification of and access to past papers

(researchers who read many papers sometimes only vaguely remember or completely

forget what they read in the past). The TogoDoc system is available for both

Windows and Mac OS X and is free. The TogoDoc Client software is available at

http://tdc.cb.k.u-tokyo.ac.jp/, and the TogoDoc server is available at

https://docman.dbcls.jp/pubmed\_recom.

DOI: 10.1371/journal.pone.0015305

PMCID: PMC3001491

PMID: 21179453 [Indexed for MEDLINE]

1825. PLoS One. 2010 Dec 13;5(12):e14286. doi: 10.1371/journal.pone.0014286.

Accurate protein structure annotation through competitive diffusion of enzymatic

functions over a network of local evolutionary similarities.

Venner E(1), Lisewski AM, Erdin S, Ward RM, Amin SR, Lichtarge O.

Author information:

(1)Department of Molecular and Human Genetics, Baylor College of Medicine,

Houston, Texas, United States of America.

High-throughput Structural Genomics yields many new protein structures without

known molecular function. This study aims to uncover these missing annotations by

globally comparing select functional residues across the structural proteome.

First, Evolutionary Trace Annotation, or ETA, identifies which proteins have

local evolutionary and structural features in common; next, these proteins are

linked together into a proteomic network of ETA similarities; then, starting from

proteins with known functions, competing functional labels diffuse link-by-link

over the entire network. Every node is thus assigned a likelihood z-score for

every function, and the most significant one at each node wins and defines its

annotation. In high-throughput controls, this competitive diffusion process

recovered enzyme activity annotations with 99% and 97% accuracy at half-coverage

for the third and fourth Enzyme Commission (EC) levels, respectively. This

corresponds to false positive rates 4-fold lower than nearest-neighbor and 5-fold

lower than sequence-based annotations. In practice, experimental validation of

the predicted carboxylesterase activity in a protein from Staphylococcus aureus

illustrated the effectiveness of this approach in the context of an increasingly

drug-resistant microbe. This study further links molecular function to a small

number of evolutionarily important residues recognizable by Evolutionary Tracing

and it points to the specificity and sensitivity of functional annotation by

competitive global network diffusion. A web server is at

http://mammoth.bcm.tmc.edu/networks.

DOI: 10.1371/journal.pone.0014286

PMCID: PMC3001439

PMID: 21179190 [Indexed for MEDLINE]

1826. BMC Genomics. 2010 Dec 2;11 Suppl 4:S21. doi: 10.1186/1471-2164-11-S4-S21.

SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction.

Wee LJ(1), Simarmata D, Kam YW, Ng LF, Tong JC.

Author information:

(1)Singapore Immunology Network, Biopolis, Singapore. lawrence@bic.nus.edu.sg

BACKGROUND: The identification of B-cell epitopes on antigens has been a subject

of intense research as the knowledge of these markers has great implications for

the development of peptide-based diagnostics, therapeutics and vaccines. As

experimental approaches are often laborious and time consuming, in silico methods

for prediction of these immunogenic regions are critical. Such efforts, however,

have been significantly hindered by high variability in the length and

composition of the epitope sequences, making naïve modeling methods difficult to

apply.

RESULTS: We analyzed two benchmark datasets and found that linear B-cell epitopes

possess distinctive residue conservation and position-specific residue

propensities which could be exploited for epitope discrimination in silico. We

developed a support vector machines (SVM) prediction model employing Bayes

Feature Extraction to predict linear B-cell epitopes of diverse lengths (12- to

20-mers). The best SVM classifier achieved an accuracy of 74.50% and AROC of 0.84

on an independent test set and was shown to outperform existing linear B-cell

epitope prediction algorithms. In addition, we applied our model to a dataset of

antigenic proteins with experimentally-verified epitopes and found it to be

generally effective for discriminating the epitopes from non-epitopes.

CONCLUSION: We developed a SVM prediction model utilizing Bayes Feature

Extraction and showed that it was effective in discriminating epitopes from

non-epitopes in benchmark datasets and annotated antigenic proteins. A web server

for predicting linear B-cell epitopes was developed and is available, together

with supplementary materials, at http://www.immunopred.org/bayesb/index.html.

DOI: 10.1186/1471-2164-11-S4-S21

PMCID: PMC3005920

PMID: 21143805 [Indexed for MEDLINE]

1827. BMC Genomics. 2010 Dec 2;11 Suppl 4:S20. doi: 10.1186/1471-2164-11-S4-S20.

Gerontome: a web-based database server for aging-related genes and analysis

pipelines.

Kwon J(1), Lee B, Chung H.

Author information:

(1)Korean BioInformation Center (KOBIC), KRIBB, Daejeon, Korea. kwon@kribb.re.kr

BACKGROUND: Aging is a complex and challenging phenomenon that requires

interdisciplinary efforts to unravel its mystery. Insight into genes relevant to

the aging process would offer the chance to delay and avoid some of deteriorative

aspects of aging through the use of preventive methods. To assist basic research

on aging, a comprehensive database and analysis platform for aging-related genes

is required.

RESULTS: We developed a web-based database server, called Gerontome that contains

aging-related gene information and user-friendly analysis pipelines. To construct

the Gerontome database, we integrated aging-related genes and their annotation

data. The aging-related genes were categorized by a set of structural terms from

Gene Ontology (GO). Analysis pipelines for promoter analysis and protein-ligand

docking were developed. The promoter analysis pipeline allows users to

investigate the age-dependent regulation of gene expression. The protein-ligand

docking pipeline provides information on the position and orientation of a ligand

in an age-related protein surface.

CONCLUSION: Gerontome can be accessed through web interfaces for querying and

browsing. The server provides comprehensive age-related gene information and

analysis pipelines. Gerontome is available free at http://gerontome.kobic.re.kr.

DOI: 10.1186/1471-2164-11-S4-S20

PMCID: PMC3005931

PMID: 21143804 [Indexed for MEDLINE]

1828. Bioinformatics. 2010 Dec 1;26(23):2986-7. doi: 10.1093/bioinformatics/btq582.

Epub 2010 Oct 17.

FILTREST3D: discrimination of structural models using restraints from

experimental data.

Gajda MJ(1), Tuszynska I, Kaczor M, Bakulina AY, Bujnicki JM.

Author information:

(1)International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4,

Warsaw, Poland.

SUMMARY: Automatic methods for macromolecular structure prediction (fold

recognition, de novo folding and docking programs) produce large sets of

alternative models. These large model sets often include many native-like

structures, which are often scored as false positives. Such native-like models

can be more easily identified based on data from experimental analyses used as

structural restraints (e.g. identification of nearby residues by cross-linking,

chemical modification, site-directed mutagenesis, deuterium exchange coupled with

mass spectrometry, etc.). We present a simple server for scoring and ranking of

models according to their agreement with user-defined restraints.

AVAILABILITY: FILTREST3D is freely available for users as a web server and

standalone software at: http://filtrest3d.genesilico.pl/

CONTACT: iamb@genesilico.pl

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq582

PMCID: PMC2982159

PMID: 20956242 [Indexed for MEDLINE]

1829. Bioinformatics. 2010 Dec 1;26(23):2997-9. doi: 10.1093/bioinformatics/btq585.

Epub 2010 Oct 14.

CellPublisher: a web platform for the intuitive visualization and sharing of

metabolic, signalling and regulatory pathways.

Flórez LA(1), Lammers CR, Michna R, Stülke J.

Author information:

(1)Department of General Microbiology, Georg-August University Göttingen,

Grisebachstr. 8, 37077 Göttingen, Germany.

SUMMARY: Systems biology relies increasingly on collaborations between several

groups with different expertise. Therefore, the systems biology community is

adopting standards that allow effective communication of concepts, as well as

transmission and processing of pathway information. The Systems Biology Graphical

Notation (SBGN) is a graphical language for biological pathways that has both a

biological as well as a computational meaning. The program CellDesigner allows

the codification of biological phenomena in an SBGN compliant form. CellPublisher

is a web server that allows the conversion of CellDesigner files to web-based

navigatable diagrams based on the user interface of Google maps. Thus,

CellPublisher complements CellDesigner by facilitating the understanding of

complex diagrams and by providing the possibility to share any CellDesigner

diagram online with collaborators and get their feedback. Due to the intuitive

interface of the online diagrams, CellPublisher serves as a basis for discovery

of novel properties of the modelled networks.

AVAILABILITY: The freely available web server and the documentation can be

accessed at: http://cellpublisher.gobics.de/. The source code and the offline

version for Microsoft Windows are freely available at

http://sourceforge.net/projects/cellpublisher/.

CONTACT: jstuelk@gwdg.de

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq585

PMID: 20947526 [Indexed for MEDLINE]

1830. Bioinformatics. 2010 Dec 1;26(23):2988-9. doi: 10.1093/bioinformatics/btq584.

Epub 2010 Oct 14.

The SMARTCyp cytochrome P450 metabolism prediction server.

Rydberg P(1), Gloriam DE, Olsen L.

Author information:

(1)Department of Medicinal Chemistry, Faculty of Pharmaceutical Sciences,

University of Copenhagen, Universitetsparken 2, DK-2100 Copenhagen, Denmark.

SUMMARY: The SMARTCyp server is the first web application for site of metabolism

prediction of cytochrome P450-mediated drug metabolism.

AVAILABILITY: The SMARTCyp server is freely available for use on the web at

www.farma.ku.dk/smartcyp where the SMARTCyp Java program and source code is also

available for download.

CONTACT: smartcyp@farma.ku.dk; lo@farma.ku.dk

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq584

PMID: 20947523 [Indexed for MEDLINE]

1831. Bioinformatics. 2010 Dec 1;26(23):2981-2. doi: 10.1093/bioinformatics/btq566.

Epub 2010 Oct 6.

PyETV: a PyMOL evolutionary trace viewer to analyze functional site predictions

in protein complexes.

Lua RC(1), Lichtarge O.

Author information:

(1)Department of Molecular and Human Genetics, Baylor College of Medicine,

Houston, TX 77030, USA.

SUMMARY: PyETV is a PyMOL plugin for viewing, analyzing and manipulating

predictions of evolutionarily important residues and sites in protein structures

and their complexes. It seamlessly captures the output of the Evolutionary Trace

server, namely ranked importance of residues, for multiple chains of a complex.

It then yields a high resolution graphical interface showing their distribution

and clustering throughout a quaternary structure, including at interfaces.

Together with other tools in the popular PyMOL viewer, PyETV thus provides a

novel tool to integrate evolutionary forces into the design of experiments

targeting the most functionally relevant sites of a protein.

AVAILABILITY: The PyETV module is written in Python. Installation instructions

and video demonstrations may be found at the URL

http://mammoth.bcm.tmc.edu/traceview/HelpDocs/PyETVHelp/pyInstructions.html.

CONTACT: lichtarge@bcm.tmc.edu.

DOI: 10.1093/bioinformatics/btq566

PMCID: PMC2982157

PMID: 20929911 [Indexed for MEDLINE]

1832. J Bioinform Comput Biol. 2010 Dec;8 Suppl 1:17-32.

BiMFG: bioinformatics tools for marine and freshwater species.

Shih TH(1), Chen CM, Wang HW, Pai TW, Chang HT.

Author information:

(1)Department of Computer Science and Engineering, National Taiwan Ocean

University, Keelung, Taiwan, Republic of China.

Biomolecule sequences and structures of land, air and water species are

determined rapidly and the data entries are unevenly distributed for different

organisms. It frequently leads to the BLAST results of homologous search

containing undesirable entries from organisms living in different environments.

To reduce irrelevant searching results, a separate database for comparative

genomics is urgently required. A comprehensive bioinformatics tool set and an

integrated database, named Bioinformatics tools for Marine and Freshwater

Genomics (BiMFG), are constructed for comparative analyses among model species

and underwater species. Novel matching techniques based on conserved motifs

and/or secondary structure elements are designed for efficiently and effectively

retrieving and aligning remote sequences through cross-species comparisons. It is

especially helpful when sequences under analysis possess low similarities and

unresolved structural information. In addition, the system provides core

techniques of multiple sequence alignment, multiple second structure profile

alignment and iteratively refined multiple structural alignments for biodiversity

analysis and verification in marine and freshwater biology. The BiMFG web server

is freely available for use at http://bimfg.cs.ntou.edu.tw/.

PMID: 21155017 [Indexed for MEDLINE]

1833. J Bioinform Comput Biol. 2010 Dec;8(6):967-80.

On comparing two structured RNA multiple alignments.

Patel V(1), Wang JT, Setia S, Verma A, Warden CD, Zhang K.

Author information:

(1)Department of Computer Science, New Jersey Institute of Technology, Newark, NJ

07102, USA. vgp22@njit.edu

We present a method, called BlockMatch, for aligning two blocks, where a block is

an RNA multiple sequence alignment with the consensus secondary structure of the

alignment in Stockholm format. The method employs a quadratic-time dynamic

programming algorithm for aligning columns and column pairs of the multiple

alignments in the blocks. Unlike many other tools that can perform pairwise

alignment of either single sequences or structures only, BlockMatch takes into

account the characteristics of all the sequences in the blocks along with their

consensus structures during the alignment process, thus being able to achieve a

high-quality alignment result. We apply BlockMatch to phylogeny reconstruction on

a set of 5S rRNA sequences taken from fifteen bacteria species. Experimental

results showed that the phylogenetic tree generated by our method is more

accurate than the tree constructed based on the widely used ClustalW tool. The

BlockMatch algorithm is implemented into a web server, accessible at

http://bioinformatics.njit.edu/blockmatch. A jar file of the program is also

available for download from the web server.

PMID: 21121021 [Indexed for MEDLINE]

1834. J Biomol Struct Dyn. 2010 Dec;28(3):415-9.

The MBLOSUM: a server for deriving mutation targets and position-specific

substitution rates.

Ma BG(1), Berezovsky IN.

Author information:

(1)Computational Biology Unit, Bergen Center for Computational Science University

of Bergen, Bergen 5008, Norway.

To facilitate mutagenesis study, it is necessary to be able to derive mutation

targets and associated substitution rates in the sequence of interest regardless

of the availability of corresponding structure. It is also important to obtain

these data depending on the specific aims of the mutation process. The MBLOSUM

server determines candidate positions for mutations and derives position-specific

substitution rates given only a protein sequence. Different sets of complete

genomes collected according to their phylogeny or specificity of environments

along with compete set of non-redundant sequences can be used in calculations

depending on the experimental task. MBLOSUM server is available at:

http://apps.cbu.uib.no/mblosum.

DOI: 10.1080/07391102.2010.10507370

PMID: 20919756 [Indexed for MEDLINE]

1835. Protein Pept Lett. 2010 Dec;17(12):1536-41.

Prediction of interaction between enzymes and small molecules in metabolic

pathways through integrating multiple classifiers.

Lu J(1), Zhu Y, Li Y, Lu W, Hu L, Niu B, Qing P, Gu L.

Author information:

(1)School of Materials Science and Engineering, Shanghai University, 149

Yan-Chang Road, Shanghai 200444, China.

Information about interactions between enzymes and small molecules is important

for understanding various metabolic bioprocesses. In this article we applied a

majority voting system to predict the interactions between enzymes and small

molecules in the metabolic pathways, by combining several classifiers including

AdaBoost, Bagging and KNN together. The advantage of such a strategy is based on

the principle that a predictor based majority voting systems usually provide more

reliable results than any single classifier. The prediction accuracies thus

obtained on a training dataset and an independent testing dataset were 82.8% and

84.8%, respectively. The prediction accuracy for the networking couples in the

independent testing dataset was 75.5%, which is about 4% higher than that

reported in a previous study. The web-server for the prediction method presented

in this paper is available at http://chemdata.shu.edu.cn/small-enz.

PMID: 20937036 [Indexed for MEDLINE]

1836. BMC Bioinformatics. 2010 Nov 26;11:579. doi: 10.1186/1471-2105-11-579.

webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment

browser.

Löytynoja A(1), Goldman N.

Author information:

(1)EMBL-European Bioinformatics Institute, Hinxton, Cambridgeshire, UK.

ari@ebi.ac.uk

BACKGROUND: Phylogeny-aware progressive alignment has been found to perform well

in phylogenetic alignment benchmarks and to produce superior alignments for the

inference of selection on codon sequences. Its implementation in the PRANK

alignment program package also allows modelling of complex evolutionary processes

and inference of posterior probabilities for sequence sites evolving under each

distinct scenario, either simultaneously with the alignment of sequences or as a

post-processing step for an existing alignment. This has led to software with

many advanced features, and users may find it difficult to generate optimal

alignments, visualise the full information in their alignment results, or

post-process these results, e.g. by objectively selecting subsets of alignment

sites.

RESULTS: We have created a web server called webPRANK that provides an

easy-to-use interface to the PRANK phylogeny-aware alignment algorithm. The

webPRANK server supports the alignment of DNA, protein and codon sequences as

well as protein-translated alignment of cDNAs, and includes built-in structure

models for the alignment of genomic sequences. The resulting alignments can be

exported in various formats widely used in evolutionary sequence analyses. The

webPRANK server also includes a powerful web-based alignment browser for the

visualisation and post-processing of the results in the context of a cladogram

relating the sequences, allowing (e.g.) removal of alignment columns with low

posterior reliability. In addition to de novo alignments, webPRANK can be used

for the inference of ancestral sequences with phylogenetically realistic gap

patterns, and for the annotation and post-processing of existing alignments. The

webPRANK server is freely available on the web at http://tinyurl.com/webprank .

CONCLUSIONS: The webPRANK server incorporates phylogeny-aware multiple sequence

alignment, visualisation and post-processing in an easy-to-use web interface. It

widens the user base of phylogeny-aware multiple sequence alignment and allows

the performance of all alignment-related activity for small sequence analysis

projects using only a standard web browser.

DOI: 10.1186/1471-2105-11-579

PMCID: PMC3009689

PMID: 21110866 [Indexed for MEDLINE]

1837. BMC Res Notes. 2010 Nov 18;3:312. doi: 10.1186/1756-0500-3-312.

MetNetAPI: A flexible method to access and manipulate biological network data

from MetNet.

Sucaet Y(1), Wurtele ES.

Author information:

(1)Department of Genetics, Development and Cell Biology, Iowa State University,

Ames, IA 50011, USA. mash@iastate.edu.

BACKGROUND: Convenient programmatic access to different biological databases

allows automated integration of scientific knowledge. Many databases support a

function to download files or data snapshots, or a webservice that offers "live"

data. However, the functionality that a database offers cannot be represented in

a static data download file, and webservices may consume considerable

computational resources from the host server.

RESULTS: MetNetAPI is a versatile Application Programming Interface (API) to the

MetNetDB database. It abstracts, captures and retains operations away from a

biological network repository and website. A range of database functions,

previously only available online, can be immediately (and independently from the

website) applied to a dataset of interest. Data is available in four layers:

molecular entities, localized entities (linked to a specific organelle),

interactions, and pathways. Navigation between these layers is intuitive (e.g.

one can request the molecular entities in a pathway, as well as request in what

pathways a specific entity participates). Data retrieval can be customized:

Network objects allow the construction of new and integration of existing

pathways and interactions, which can be uploaded back to our server. In contrast

to webservices, the computational demand on the host server is limited to

processing data-related queries only.

CONCLUSIONS: An API provides several advantages to a systems biology software

platform. MetNetAPI illustrates an interface with a central repository of data

that represents the complex interrelationships of a metabolic and regulatory

network. As an alternative to data-dumps and webservices, it allows access to a

current and "live" database and exposes analytical functions to application

developers. Yet it only requires limited resources on the server-side (thin

server/fat client setup). The API is available for Java, Microsoft.NET and R

programming environments and offers flexible query and broad data- retrieval

methods. Data retrieval can be customized to client needs and the API offers a

framework to construct and manipulate user-defined networks. The design

principles can be used as a template to build programmable interfaces for other

biological databases. The API software and tutorials are available at

http://www.metnetonline.org/api.

DOI: 10.1186/1756-0500-3-312

PMCID: PMC2998519

PMID: 21083943

1838. Bioinformatics. 2010 Nov 15;26(22):2904-5. doi: 10.1093/bioinformatics/btq547.

Epub 2010 Oct 12.

PICMI: mapping point mutations on genomes.

Le Pera L(1), Marcatili P, Tramontano A.

Author information:

(1)Department of Biochemical Sciences, Sapienza University of Rome, P.le A. Moro,

5-00185 Rome, Italy.

MOTIVATION: Several international collaborations and local projects are producing

extensive catalogues of genomic variations that are supplementing existing

collections such as the OMIM catalogue. The flood of this type of data will keep

increasing and, especially, it will be relevant to a wider user base, including

not only molecular biologists, geneticists and bioinformaticians, but also

clinical researchers. Mapping the observed variations, sometimes only described

at the amino acid level, on a genome, identifying whether they affect a gene

and-if so-whether they also affect different isoforms of the same gene, is a time

consuming and often frustrating task.

RESULTS: The PICMI server is an easy to use tool for quickly mapping one or more

amino acid or nucleotide variations on a genome and its products, including

alternatively spliced isoforms.

AVAILABILITY: The server is available at www.biocomputing.it/picmi.

DOI: 10.1093/bioinformatics/btq547

PMCID: PMC2971578

PMID: 20940168 [Indexed for MEDLINE]

1839. Bioinformatics. 2010 Nov 15;26(22):2914-5. doi: 10.1093/bioinformatics/btq549.

Epub 2010 Sep 27.

GLOOME: gain loss mapping engine.

Cohen O(1), Ashkenazy H, Belinky F, Huchon D, Pupko T.

Author information:

(1)Department of Cell Research and Immunology, Tel-Aviv University, Tel Aviv

69978, Israel.

The evolutionary analysis of presence and absence profiles (phyletic patterns) is

widely used in biology. It is assumed that the observed phyletic pattern is the

result of gain and loss dynamics along a phylogenetic tree. Examples of

characters that are represented by phyletic patterns include restriction sites,

gene families, introns and indels, to name a few. Here, we present a

user-friendly web server that accurately infers branch-specific and site-specific

gain and loss events. The novel inference methodology is based on a stochastic

mapping approach utilizing models that reliably capture the underlying

evolutionary processes. A variety of features are available including the ability

to analyze the data with various evolutionary models, to infer gain and loss

events using either stochastic mapping or maximum parsimony, and to estimate gain

and loss rates for each character analyzed.AVAILABILITY: Freely available for use

at http://gloome.tau.ac.il/.

DOI: 10.1093/bioinformatics/btq549

PMID: 20876605 [Indexed for MEDLINE]

1840. Bioinformatics. 2010 Nov 15;26(22):2900-1. doi: 10.1093/bioinformatics/btq545.

Epub 2010 Sep 24.

SylArray: a web server for automated detection of miRNA effects from expression

data.

Bartonicek N(1), Enright AJ.

Author information:

(1)EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

A useful step for understanding the function of microRNAs (miRNA) or siRNAs is

the detection of their effects on genome-wide expression profiles. Typically,

approaches look for enrichment of words in the 3(')UTR sequences of the most

deregulated genes. A number of tools are available for this purpose, but they

require either in-depth computational knowledge, filtered 3(')UTR sequences for

the genome of interest, or a set of genes acquired through an arbitrary

expression cutoff. To this end, we have developed SylArray; a web-based resource

designed for the analysis of large-scale expression datasets. It simply requires

the user to submit a sorted list of genes from an expression experiment. SylArray

utilizes curated sets of 3(')UTRs to attach sequences to these genes and then

applies the Sylamer algorithm for detection of miRNA or siRNA signatures in those

sequences. An intuitive system for visualization and interpretation of the small

RNA signatures is included.AVAILABILITY: SylArray is written in Perl-CGI, Perl

and Java and also uses the R statistical package. The source-code, database and

web resource are freely available under GNU Public License (GPL). The web server

is freely accessible at http://www.ebi.ac.uk/enright/sylarray.

DOI: 10.1093/bioinformatics/btq545

PMID: 20871108 [Indexed for MEDLINE]

1841. Bioinformatics. 2010 Nov 15;26(22):2916-7. doi: 10.1093/bioinformatics/btq537.

Epub 2010 Sep 21.

MOBI: a web server to define and visualize structural mobility in NMR protein

ensembles.

Martin AJ(1), Walsh I, Tosatto SC.

Author information:

(1)Department of Biology, University of Padova, viale G. Colombo 3, 35131 Padova,

Italy.

MOTIVATION: MOBI is a web server for the identification of structurally mobile

regions in NMR protein ensembles. It provides a binary mobility definition that

is analogous to the commonly used definition of intrinsic disorder in X-ray

crystallographic structures. At least three different use cases can be envisaged:

(i) visualization of NMR mobility for structural analysis; (ii) definition of

regions for reliable comparative modelling in protein structure prediction and

(iii) definition of mobility in analogy to intrinsic disorder. MOBI uses

structural superposition and local conformational differences to derive a robust

binary mobility definition that is in excellent agreement with the manually

curated definition used in the CASP8 experiment for intrinsic disorder in NMR

structure. The output includes mobility-coloured PDB files, mobility plots and a

FASTA formatted sequence file summarizing the mobility results.

AVAILABILITY: The MOBI server and supplementary methods are available for

non-commercial use at URL: http://protein.bio.unipd.it/mobi/.

DOI: 10.1093/bioinformatics/btq537

PMID: 20861031 [Indexed for MEDLINE]

1842. Bioinformatics. 2010 Nov 15;26(22):2897-9. doi: 10.1093/bioinformatics/btq540.

Epub 2010 Sep 21.

FastPval: a fast and memory efficient program to calculate very low P-values from

empirical distribution.

Li MJ(1), Sham PC, Wang J.

Author information:

(1)Department of Biochemistry, LKS Faculty of Medicine, The University of Hong

Kong, 21 Sassoon Rd, Pokfulam, Hong Kong SAR, China.

MOTIVATION: Resampling methods, such as permutation and bootstrap, have been

widely used to generate an empirical distribution for assessing the statistical

significance of a measurement. However, to obtain a very low P-value, a large

size of resampling is required, where computing speed, memory and storage

consumption become bottlenecks, and sometimes become impossible, even on a

computer cluster.

RESULTS: We have developed a multiple stage P-value calculating program called

FastPval that can efficiently calculate very low (up to 10(-9)) P-values from a

large number of resampled measurements. With only two input files and a few

parameter settings from the users, the program can compute P-values from

empirical distribution very efficiently, even on a personal computer. When tested

on the order of 10(9) resampled data, our method only uses 52.94% the time used

by the conventional method, implemented by standard quicksort and binary search

algorithms, and consumes only 0.11% of the memory and storage. Furthermore, our

method can be applied to extra large datasets that the conventional method fails

to calculate. The accuracy of the method was tested on data generated from

Normal, Poison and Gumbel distributions and was found to be no different from the

exact ranking approach.

AVAILABILITY: The FastPval executable file, the java GUI and source code, and the

java web start server with example data and introduction, are available at

http://wanglab.hku.hk/pvalue.

DOI: 10.1093/bioinformatics/btq540

PMCID: PMC2971576

PMID: 20861029 [Indexed for MEDLINE]

1843. PLoS One. 2010 Nov 15;5(11):e13934. doi: 10.1371/journal.pone.0013934.

High sensitivity TSS prediction: estimates of locations where TSS cannot occur.

Schaefer U(1), Kodzius R, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bajic VB.

Author information:

(1)Computational Bioscience Research Center (CBRC), King Abdullah University of

Science and Technology, Thuwal, Kingdom of Saudi Arabia.

BACKGROUND: Although transcription in mammalian genomes can initiate from various

genomic positions (e.g., 3'UTR, coding exons, etc.), most locations on genomes

are not prone to transcription initiation. It is of practical and theoretical

interest to be able to estimate such collections of non-TSS locations (NTLs). The

identification of large portions of NTLs can contribute to better focusing the

search for TSS locations and thus contribute to promoter and gene finding. It can

help in the assessment of 5' completeness of expressed sequences, contribute to

more successful experimental designs, as well as more accurate gene annotation.

METHODOLOGY: Using comprehensive collections of Cap Analysis of Gene Expression

(CAGE) and other transcript data from mouse and human genomes, we developed a

methodology that allows us, by performing computational TSS prediction with very

high sensitivity, to annotate, with a high accuracy in a strand specific manner,

locations of mammalian genomes that are highly unlikely to harbor transcription

start sites (TSSs). The properties of the immediate genomic neighborhood of

98,682 accurately determined mouse and 113,814 human TSSs are used to determine

features that distinguish genomic transcription initiation locations from those

that are not likely to initiate transcription. In our algorithm we utilize

various constraining properties of features identified in the upstream and

downstream regions around TSSs, as well as statistical analyses of these

surrounding regions.

CONCLUSIONS: Our analysis of human chromosomes 4, 21 and 22 estimates ∼46%, ∼41%

and ∼27% of these chromosomes, respectively, as being NTLs. This suggests that on

average more than 40% of the human genome can be expected to be highly unlikely

to initiate transcription. Our method represents the first one that utilizes

high-sensitivity TSS prediction to identify, with high accuracy, large portions

of mammalian genomes as NTLs. The server with our algorithm implemented is

available at http://cbrc.kaust.edu.sa/ddm/.

DOI: 10.1371/journal.pone.0013934

PMCID: PMC2981523

PMID: 21085627 [Indexed for MEDLINE]

1844. BMC Bioinformatics. 2010 Nov 11;11:555. doi: 10.1186/1471-2105-11-555.

The LabelHash algorithm for substructure matching.

Moll M(1), Bryant DH, Kavraki LE.

Author information:

(1)Department of Computer Science, Rice University, Houston, TX 77005, USA.

mmoll@rice.edu

BACKGROUND: There is an increasing number of proteins with known structure but

unknown function. Determining their function would have a significant impact on

understanding diseases and designing new therapeutics. However, experimental

protein function determination is expensive and very time-consuming.

Computational methods can facilitate function determination by identifying

proteins that have high structural and chemical similarity.

RESULTS: We present LabelHash, a novel algorithm for matching substructural

motifs to large collections of protein structures. The algorithm consists of two

phases. In the first phase the proteins are preprocessed in a fashion that allows

for instant lookup of partial matches to any motif. In the second phase, partial

matches for a given motif are expanded to complete matches. The general

applicability of the algorithm is demonstrated with three different case studies.

First, we show that we can accurately identify members of the enolase superfamily

with a single motif. Next, we demonstrate how LabelHash can complement SOIPPA, an

algorithm for motif identification and pairwise substructure alignment. Finally,

a large collection of Catalytic Site Atlas motifs is used to benchmark the

performance of the algorithm. LabelHash runs very efficiently in parallel;

matching a motif against all proteins in the 95% sequence identity filtered

non-redundant Protein Data Bank typically takes no more than a few minutes. The

LabelHash algorithm is available through a web server and as a suite of

standalone programs at http://labelhash.kavrakilab.org. The output of the

LabelHash algorithm can be further analyzed with Chimera through a plugin that we

developed for this purpose.

CONCLUSIONS: LabelHash is an efficient, versatile algorithm for large-scale

substructure matching. When LabelHash is running in parallel, motifs can

typically be matched against the entire PDB on the order of minutes. The

algorithm is able to identify functional homologs beyond the twilight zone of

sequence identity and even beyond fold similarity. The three case studies

presented in this paper illustrate the versatility of the algorithm.

DOI: 10.1186/1471-2105-11-555

PMCID: PMC2996407

PMID: 21070651 [Indexed for MEDLINE]

1845. J Theor Biol. 2010 Nov 7;267(1):29-34. doi: 10.1016/j.jtbi.2010.08.007. Epub 2010

Aug 7.

2D-MH: A web-server for generating graphic representation of protein sequences

based on the physicochemical properties of their constituent amino acids.

Wu ZC(1), Xiao X, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333046,

China.

Introduction of graphic representation for biological sequences can provide

intuitive overall pictures as well as useful insights for performing large-scale

analysis. Here, a new two-dimensional graph, called "2D-MH", is proposed to

represent protein sequences. It is formed by incorporating the information of the

side-chain mass of each of the constituent amino acids and its hydrophobicity.

The graphic curve thus generated is featured by (1) an one-to-one correspondence

relation without circuit or degeneracy, (2) better reflecting the innate

structure of the protein sequence, (3) clear visibility in displaying the

similarity of protein sequences, (4) more sensitive for the mutation sites

important for drug targeting, and (5) being able to be used as a metric for the

"evolutionary distance" of a protein from one species to the other. It is

anticipated that the presented graphic method may become a useful vehicle for

large-scale analysis of the avalanche of protein sequences generated in the

post-genomic age. As a web-server, 2D-MH is freely accessible at

http://icpr.jci.jx.cn/bioinfo/pplot/2D-MH, by which one can easily generate the

two-dimensional graphs for any number of protein sequences and compare the

evolutionary distances between them.

Copyright © 2010 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2010.08.007

PMID: 20696175 [Indexed for MEDLINE]

1846. J Theor Biol. 2010 Nov 7;267(1):1-6. doi: 10.1016/j.jtbi.2010.08.001. Epub 2010

Aug 5.

SecretP: identifying bacterial secreted proteins by fusing new features into

Chou's pseudo-amino acid composition.

Yu L(1), Guo Y, Li Y, Li G, Li M, Luo J, Xiong W, Qin W.

Author information:

(1)College of Chemistry, Sichuan University, Chengdu 610064, PR China.

Protein secretion plays an important role in bacterial lifestyles. Secreted

proteins are crucial for bacterial pathogenesis by making bacteria interact with

their environments, particularly delivering pathogenic and symbiotic bacteria

into their eukaryotic hosts. Therefore, identification of bacterial secreted

proteins becomes an important process for the study of various diseases and the

corresponding drugs. In this paper, fusing several new features into Chou's

pseudo-amino acid composition (PseAAC), two support vector machine (SVM)-based

ternary classifiers are developed to predict secreted proteins of Gram-negative

and Gram-positive bacteria. For the two types of bacteria, the high accuracy of

94.03% and 94.36% are obtained in distinguishing classically secreted,

non-classically secreted and non-secreted proteins by our method. In order to

compare the practical ability of our method in identifying bacterial secreted

proteins with those of six published methods, proteins in Escherichia coli and

Bacillus subtilis are collected to construct the test sets of Gram-negative and

Gram-positive bacteria, and the prediction results of our method are comparable

to those of existing methods. When performed on two public independent data sets

for predicting NCSPs, it also yields satisfactory results for Gram-negative

bacterial proteins. The prediction server SecretP can be accessed at

http://cic.scu.edu.cn/bioinformatics/secretPV2/index.htm.

Copyright © 2010 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2010.08.001

PMID: 20691704 [Indexed for MEDLINE]

1847. BMC Bioinformatics. 2010 Nov 2;11:543. doi: 10.1186/1471-2105-11-543.

Predicting success of oligomerized pool engineering (OPEN) for zinc finger target

site sequences.

Sander JD(1), Reyon D, Maeder ML, Foley JE, Thibodeau-Beganny S, Li X, Regan MR,

Dahlborg EJ, Goodwin MJ, Fu F, Voytas DF, Joung JK, Dobbs D.

Author information:

(1)Molecular Pathology Unit, Center for Cancer Research, and Center for

Computational and Integrative Biology, Massachusetts General Hospital,

Charlestown, MA 02129, USA. jsander@partners.org

BACKGROUND: Precise and efficient methods for gene targeting are critical for

detailed functional analysis of genomes and regulatory networks and for

potentially improving the efficacy and safety of gene therapies. Oligomerized

Pool ENgineering (OPEN) is a recently developed method for engineering C2H2 zinc

finger proteins (ZFPs) designed to bind specific DNA sequences with high affinity

and specificity in vivo. Because generation of ZFPs using OPEN requires

considerable effort, a computational method for identifying the sites in any

given gene that are most likely to be successfully targeted by this method is

desirable.

RESULTS: Analysis of the base composition of experimentally validated ZFP target

sites identified important constraints on the DNA sequence space that can be

effectively targeted using OPEN. Using alternate encodings to represent ZFP

target sites, we implemented Naïve Bayes and Support Vector Machine classifiers

capable of distinguishing "active" targets, i.e., ZFP binding sites that can be

targeted with a high rate of success, from those that are "inactive" or poor

targets for ZFPs generated using current OPEN technologies. When evaluated using

leave-one-out cross-validation on a dataset of 135 experimentally validated ZFP

target sites, the best Naïve Bayes classifier, designated ZiFOpT, achieved

overall accuracy of 87% and specificity+ of 90%, with an ROC AUC of 0.89. When

challenged with a completely independent test set of 140 newly validated ZFP

target sites, ZiFOpT performance was comparable in terms of overall accuracy

(88%) and specificity+ (92%), but with reduced ROC AUC (0.77). Users can rank

potentially active ZFP target sites using a confidence score derived from the

posterior probability returned by ZiFOpT.

CONCLUSION: ZiFOpT, a machine learning classifier trained to identify DNA

sequences amenable for targeting by OPEN-generated zinc finger arrays, can guide

users to target sites that are most likely to function successfully in vivo,

substantially reducing the experimental effort required. ZiFOpT is freely

available and incorporated in the Zinc Finger Targeter web server

(http://bindr.gdcb.iastate.edu/ZiFiT).

DOI: 10.1186/1471-2105-11-543

PMCID: PMC3098093

PMID: 21044337 [Indexed for MEDLINE]

1848. BMC Genomics. 2010 Nov 2;11 Suppl 2:S5. doi: 10.1186/1471-2164-11-S2-S5.

Sequence feature-based prediction of protein stability changes upon amino acid

substitutions.

Teng S(1), Srivastava AK, Wang L.

Author information:

(1)Department of Genetics and Biochemistry, Clemson University, Clemson, SC

29634, USA. steng@clemson.edu

BACKGROUND: Protein destabilization is a common mechanism by which amino acid

substitutions cause human diseases. Although several machine learning methods

have been reported for predicting protein stability changes upon amino acid

substitutions, the previous studies did not utilize relevant sequence features

representing biological knowledge for classifier construction.

RESULTS: In this study, a new machine learning method has been developed for

sequence feature-based prediction of protein stability changes upon amino acid

substitutions. Support vector machines were trained with data from experimental

studies on the free energy change of protein stability upon mutations. To

construct accurate classifiers, twenty sequence features were examined for input

vector encoding. It was shown that classifier performance varied significantly by

using different sequence features. The most accurate classifier in this study was

constructed using a combination of six sequence features. This classifier

achieved an overall accuracy of 84.59% with 70.29% sensitivity and 90.98%

specificity.

CONCLUSIONS: Relevant sequence features can be used to accurately predict protein

stability changes upon amino acid substitutions. Predictive results at this level

of accuracy may provide useful information to distinguish between deleterious and

tolerant alterations in disease candidate genes. To make the classifier

accessible to the genetics research community, we have developed a new web

server, called MuStab (http://bioinfo.ggc.org/mustab/).

DOI: 10.1186/1471-2164-11-S2-S5

PMCID: PMC2975416

PMID: 21047386 [Indexed for MEDLINE]

1849. Bioinformatics. 2010 Nov 1;26(21):2689-97. doi: 10.1093/bioinformatics/btq506.

Epub 2010 Oct 6.

R3D Align: global pairwise alignment of RNA 3D structures using local

superpositions.

Rahrig RR(1), Leontis NB, Zirbel CL.

Author information:

(1)Department of Mathematics and Statistics at Ohio Northern University, Ada, OH

45810, USA. r-rahrig@onu.edu

MOTIVATION: Comparing 3D structures of homologous RNA molecules yields

information about sequence and structural variability. To compare large RNA 3D

structures, accurate automatic comparison tools are needed. In this article, we

introduce a new algorithm and web server to align large homologous RNA structures

nucleotide by nucleotide using local superpositions that accommodate the

flexibility of RNA molecules. Local alignments are merged to form a global

alignment by employing a maximum clique algorithm on a specially defined graph

that we call the 'local alignment' graph.

RESULTS: The algorithm is implemented in a program suite and web server called

'R3D Align'. The R3D Align alignment of homologous 3D structures of 5S, 16S and

23S rRNA was compared to a high-quality hand alignment. A full comparison of the

16S alignment with the other state-of-the-art methods is also provided. The R3D

Align program suite includes new diagnostic tools for the structural evaluation

of RNA alignments. The R3D Align alignments were compared to those produced by

other programs and were found to be the most accurate, in comparison with a high

quality hand-crafted alignment and in conjunction with a series of other

diagnostics presented. The number of aligned base pairs as well as measures of

geometric similarity are used to evaluate the accuracy of the alignments.

AVAILABILITY: R3D Align is freely available through a web server

http://rna.bgsu.edu/R3DAlign. The MATLAB source code of the program suite is also

freely available for download at that location.

DOI: 10.1093/bioinformatics/btq506

PMCID: PMC3465099

PMID: 20929913 [Indexed for MEDLINE]

1850. Bioinformatics. 2010 Nov 1;26(21):2784-5. doi: 10.1093/bioinformatics/btq504.

Epub 2010 Sep 24.

SimiCon: a web tool for protein-ligand model comparison through calculation of

equivalent atomic contacts.

Rueda M(1), Katritch V, Raush E, Abagyan R.

Author information:

(1)UCSD Skaggs School of Pharmacy and Pharmaceutical Sciences, University of

California-San Diego, La Jolla, CA 92093, USA.

SUMMARY: SimiCon is a web server designed for an automated identification of

equivalent protein-ligand atomic contacts in different conformational models of a

complex. The contacts are computed with internal coordinate mechanics (ICM)

software with respect to molecular symmetry and the results are shown in the

browser as text, tables and interactive 3D graphics. The web server can be

executed remotely without a browser to allow users to automate multiple

calculations.

AVAILABILITY: SimiCon is freely available at http://abagyan.ucsd.edu/SimiCon

DOI: 10.1093/bioinformatics/btq504

PMCID: PMC2981495

PMID: 20871105 [Indexed for MEDLINE]

1851. Hum Genet. 2010 Nov;128(5):473-9. doi: 10.1007/s00439-010-0890-8. Epub 2010 Sep

18.

The molecular basis of autosomal recessive diseases among the Arabs and Druze in

Israel.

Zlotogora J(1).

Author information:

(1)Department of Community Genetics, Public Health Services, Ministry of Health

Israel, Building 67, Sheba Medical Center, Tel Hashomer, 52621 Ramat Gan, Israel.

joelz@cc.huji.ac.il

The Israeli population mainly includes Jews, Muslim and Christian Arabs, and

Druze In the last decade, data on genetic diseases present in the population have

been systematically collected and are available online in the Israeli national

genetic database ( http://www.goldenhelix.org/server/israeli ). In the non-Jewish

population, up to 1 July 2010, the database included molecular data on six

diseases relatively frequent in the whole population: thalassemia, familial

Mediterranean fever (FMF), cystic fibrosis, deafness, phenylketonuria and

congenital adrenal hyperplasia, as well as data on 195 autosomal recessive

diseases among Muslim Israeli Arabs, 11 among the Christian Arabs and 31 among

Druze. A single mutation was characterized in 149 out of the 238 rare disorders

for which the molecular basis was known. In many diseases, mutation had never

been observed in any other population and was present in one family only

suggesting that it occurred as a de novo event. In other diseases, the mutation

was present in more than one community or even in other populations such as

Bedouins from the Arab peninsula or Christians from Lebanon. In the 89 other

disorders, more than one mutation was characterized either in the same gene or in

more than one gene. While it is probable that most of these cases represent

random events in some cases such as Bardet Biedl among the Bedouins, the reason

may be a selective advantage to the heterozygotes.

DOI: 10.1007/s00439-010-0890-8

PMID: 20852892 [Indexed for MEDLINE]

1852. Masui. 2010 Nov;59(11):1452-5.

[A central monitoring system capturing the screen of an anesthesia monitor].

[Article in Japanese]

Sakaguchi H(1), Abe E, Abe M, Anraku S, Higuchi T, Ishida T.

Author information:

(1)Department of Anesthesiology, Japanese Red Cross Kumamoto Hospital, Kumamoto

861-8520.

e have developed a flexible, economically efficient central monitoring system.

The system converts RGB analog outputs on the screen of an anesthesia monitor

display into digital video signals with TwinPact 100 (Thomson Canopus, Kobe), and

depicts them on the screen of Vidi-installed (http://www.mitzpettel.com/

software/vidi.php) personal computers (iMac, Apple, Tokyo), which serve as

terminal monitors. These PCs are monitored and administered through Apple Remote

Desktop 3 (Apple, Tokyo) on a server computer (Mac Pro, Apple, Tokyo), connected

to the LAN, in the office for anesthesiologists. As Bosco's Screen Share

(http://www.componentx.com/ScreenShare/) has been installed on computers in every

room, we can monitor their screens via a PC in another room using Firefox

(http://mozilla.jp/firefox/) and other web browsers.The system, with a screen

capturing function, was designed to comply with all monitor display of all

medical equipment manufacturers, with possible expansion to the operating rooms.

PMID: 21077323 [Indexed for MEDLINE]

1853. Mol Divers. 2010 Nov;14(4):719-29. doi: 10.1007/s11030-009-9216-y. Epub 2009 Dec

30.

Predicting miRNA's target from primary structure by the nearest neighbor

algorithm.

Lin K(1), Qian Z, Lu L, Lu L, Lai L, Gu J, Zeng Z, Li H, Cai Y.

Author information:

(1)CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for

Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai,

200031, China.

We used a machine learning method, the nearest neighbor algorithm (NNA), to learn

the relationship between miRNAs and their target proteins, generating a predictor

which can then judge whether a new miRNA-target pair is true or not. We acquired

198 positive (true) miRNA-target pairs from Tarbase and the literature, and

generated 4,888 negative (false) pairs through random combination. A 0/1 system

and the frequencies of single nucleotides and di-nucleotides were used to encode

miRNAs into vectors while various physicochemical parameters were used to encode

the targets. The NNA was then applied, learning from these data to produce a

predictor. We implemented minimum redundancy maximum relevance (mRMR) and

properties forward selection (PFS) to reduce the redundancy of our encoding

system, obtaining 91 most efficient properties. Finally, via the Jackknife

cross-validation test, we got a positive accuracy of 69.2% and an overall

accuracy of 96.0% with all the 253 properties. Besides, we got a positive

accuracy of 83.8% and an overall accuracy of 97.2% with the 91 most efficient

properties. A web-server for predictions is also made available at

http://app3.biosino.org:8080/miRTP/index.jsp.

DOI: 10.1007/s11030-009-9216-y

PMID: 20041294 [Indexed for MEDLINE]

1854. BMC Bioinformatics. 2010 Oct 28;11:535. doi: 10.1186/1471-2105-11-535.

Optimizing structural modeling for a specific protein scaffold: knottins or

inhibitor cystine knots.

Gracy J(1), Chiche L.

Author information:

(1)CNRS, UMR5048, Université Montpellier 1 et 2, Centre de Biochimie Structurale,

34090 Montpellier, France. Jerome.Gracy@cbs.cnrs.fr

BACKGROUND: Knottins are small, diverse and stable proteins with important drug

design potential. They can be classified in 30 families which cover a wide range

of sequences (1621 sequenced), three-dimensional structures (155 solved) and

functions (> 10). Inter knottin similarity lies mainly between 15% and 40%

sequence identity and 1.5 to 4.5 Å backbone deviations although they all share a

tightly knotted disulfide core. This important variability is likely to arise

from the highly diverse loops which connect the successive knotted cysteines. The

prediction of structural models for all knottin sequences would open new

directions for the analysis of interaction sites and to provide a better

understanding of the structural and functional organization of proteins sharing

this scaffold.

RESULTS: We have designed an automated modeling procedure for predicting the

three-dimensionnal structure of knottins. The different steps of the homology

modeling pipeline were carefully optimized relatively to a test set of knottins

with known structures: template selection and alignment, extraction of structural

constraints and model building, model evaluation and refinement. After

optimization, the accuracy of predicted models was shown to lie between 1.50 and

1.96 Å from native structures at 50% and 10% maximum sequence identity levels,

respectively. These average model deviations represent an improvement varying

between 0.74 and 1.17 Å over a basic homology modeling derived from a unique

template. A database of 1621 structural models for all known knottin sequences

was generated and is freely accessible from our web server at

http://knottin.cbs.cnrs.fr. Models can also be interactively constructed from any

knottin sequence using the structure prediction module Knoter1D3D available from

our protein analysis toolkit PAT at http://pat.cbs.cnrs.fr.

CONCLUSIONS: This work explores different directions for a systematic homology

modeling of a diverse family of protein sequences. In particular, we have shown

that the accuracy of the models constructed at a low level of sequence identity

can be improved by 1) a careful optimization of the modeling procedure, 2) the

combination of multiple structural templates and 3) the use of conserved

structural features as modeling restraints.

DOI: 10.1186/1471-2105-11-535

PMCID: PMC2984590

PMID: 21029427 [Indexed for MEDLINE]

1855. BMC Bioinformatics. 2010 Oct 27;11:533. doi: 10.1186/1471-2105-11-533.

Integrated prediction of one-dimensional structural features and their

relationships with conformational flexibility in helical membrane proteins.

Ahmad S(1), Singh YH, Paudel Y, Mori T, Sugita Y, Mizuguchi K.

Author information:

(1)National Institute of Biomedical Innovation, 7-6-8 Saito-asagi, Ibaraki, Osaka

567 0085, Japan. shandar@nibio.go.jp

BACKGROUND: Many structural properties such as solvent accessibility, dihedral

angles and helix-helix contacts can be assigned to each residue in a membrane

protein. Independent studies exist on the analysis and sequence-based prediction

of some of these so-called one-dimensional features. However, there is little

explanation of why certain residues are predicted in a wrong structural class or

with large errors in the absolute values of these features. On the other hand,

membrane proteins undergo conformational changes to allow transport as well as

ligand binding. These conformational changes often occur via residues that are

inherently flexible and hence, predicting fluctuations in residue positions is of

great significance.

RESULTS: We performed a statistical analysis of common patterns among selected

one-dimensional equilibrium structural features (ESFs) and developed a method for

simultaneously predicting all of these features using an integrated system. Our

results show that the prediction performance can be improved if multiple

structural features are trained in an integrated model, compared to the current

practice of developing individual models. In particular, the performance of the

solvent accessibility and bend-angle prediction improved in this way. The

well-performing bend-angle prediction can be used to predict helical positions

with severe kinks at a modest success rate. Further, we showed that single-chain

conformational dynamics, measured by B-factors derived from normal mode analysis,

could be predicted from observed and predicted ESFs with good accuracy. A web

server was developed (http://tardis.nibio.go.jp/netasa/htmone/) for predicting

the one-dimensional ESFs from sequence information and analyzing the differences

between the predicted and observed values of the ESFs.

CONCLUSIONS: The prediction performance of the integrated model is significantly

better than that of the models performing the task separately for each feature

for the solvent accessibility and bend-angle predictions. The predictability of

the features also plays a role in determining flexible positions. Although the

dynamics studied here concerns local atomic fluctuations, a similar analysis in

terms of global structural features will be helpful in predicting large-scale

conformational changes, for which work is in progress.

DOI: 10.1186/1471-2105-11-533

PMCID: PMC3247134

PMID: 20977780 [Indexed for MEDLINE]

1856. J Theor Biol. 2010 Oct 21;266(4):560-8. doi: 10.1016/j.jtbi.2010.07.026. Epub

2010 Jul 23.

Knowledge-based computational mutagenesis for predicting the disease potential of

human non-synonymous single nucleotide polymorphisms.

Masso M(1), Vaisman II.

Author information:

(1)Laboratory for Structural Bioinformatics, Department of Bioinformatics and

Computational Biology, George Mason University, 10900 University Blvd MS 5B3,

Manassas, VA 20110, USA. mmasso@gmu.edu

Certain genetic variations in the human population are associated with heritable

diseases, and single nucleotide polymorphisms (SNPs) represent the most common

form of such differences in DNA sequence. In particular, substantial interest

exists in determining whether a non-synonymous SNP (nsSNP), leading to a single

residue replacement in the translated protein product, is neutral or

disease-related. The nature of protein structure-function relationships suggests

that nsSNP effects, either benign or leading to aberrant protein function

possibly associated with disease, are dependent on relative structural changes

introduced upon mutation. In this study, we characterize a representative

sampling of 1790 documented neutral and disease-related human nsSNPs mapped to

243 diverse human protein structures, by quantifying environmental perturbations

in the associated proteins with the use of a computational mutagenesis

methodology that relies on a four-body, knowledge-based, statistical contact

potential. These structural change data are used as attributes to generate a

vector representation for each nsSNP, in combination with additional features

reflecting sequence and structure of the corresponding protein. A trained model

based on the random forest supervised classification algorithm achieves 76%

cross-validation accuracy. Our classifier performs at least as well as other

methods that use significantly larger datasets of nsSNPs for model training, and

the novelty of our attributes differentiates the model as an orthogonal approach

that can be utilized in conjunction with other techniques. A dedicated server for

obtaining predictions, as well as supporting datasets and documentation, is

available at http://proteins.gmu.edu/automute.

Copyright © 2010 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2010.07.026

PMID: 20655929 [Indexed for MEDLINE]

1857. Evol Bioinform Online. 2010 Oct 20;6:143-9. doi: 10.4137/EBO.S5861.

EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics.

Barker MS(1), Dlugosch KM, Dinh L, Challa RS, Kane NC, King MG, Rieseberg LH.

Author information:

(1)The Biodiversity Research Centre and Department of Botany, University of

British Columbia, Vancouver, BC V6T 1Z4, Canada.

Recent increases in the production of genomic data are yielding new opportunities

and challenges for biologists. Among the chief problems posed by next-generation

sequencing are assembly and analyses of these large data sets. Here we present an

online server, http://EvoPipes.net, that provides access to a wide range of tools

for bioinformatic analyses of genomic data oriented for ecological and

evolutionary biologists. The EvoPipes.net server includes a basic tool kit for

analyses of genomic data including a next-generation sequence cleaning pipeline

(SnoWhite), scaffolded assembly software (SCARF), a reciprocal best-blast hit

ortholog pipeline (RBH Orthologs), a pipeline for reference protein-based

translation and identification of reading frame in transcriptome and genomic DNA

(TransPipe), a pipeline to identify gene families and summarize the history of

gene duplications (DupPipe), and a tool for developing SSRs or microsatellites

from a transcriptome or genomic coding sequence collection (findSSR).

EvoPipes.net also provides links to other software developed for evolutionary and

ecological genomics, including chromEvol and NU-IN, as well as a forum for

discussions of issues relating to genomic analyses and interpretation of results.

Overall, these applications provide a basic bioinformatic tool kit that will

enable ecologists and evolutionary biologists with relatively little experience

and computational resources to take advantage of the opportunities provided by

next-generation sequencing in their systems.

DOI: 10.4137/EBO.S5861

PMCID: PMC2978936

PMID: 21079755

1858. Immunome Res. 2010 Oct 20;6:6. doi: 10.1186/1745-7580-6-6.

Identification of conformational B-cell Epitopes in an antigen from its primary

sequence.

Ansari HR(1), Raghava GP.

Author information:

(1)Bioinformatics Center, Institute of Microbial Technology, Sector 39-A,

Chandigarh, India. raghava@imtech.res.in.

BACKGROUND: One of the major challenges in the field of vaccine design is to

predict conformational B-cell epitopes in an antigen. In the past, several

methods have been developed for predicting conformational B-cell epitopes in an

antigen from its tertiary structure. This is the first attempt in this area to

predict conformational B-cell epitope in an antigen from its amino acid sequence.

RESULTS: All Support vector machine (SVM) models were trained and tested on 187

non-redundant protein chains consisting of 2261 antibody interacting residues of

B-cell epitopes. Models have been developed using binary profile of pattern (BPP)

and physiochemical profile of patterns (PPP) and achieved a maximum MCC of 0.22

and 0.17 respectively. In this study, for the first time SVM model has been

developed using composition profile of patterns (CPP) and achieved a maximum MCC

of 0.73 with accuracy 86.59%. We compare our CPP based model with existing

structure based methods and observed that our sequence based model is as good as

structure based methods.

CONCLUSION: This study demonstrates that prediction of conformational B-cell

epitope in an antigen is possible from is primary sequence. This study will be

very useful in predicting conformational B-cell epitopes in antigens whose

tertiary structures are not available. A web server CBTOPE has been developed for

predicting B-cell epitope http://www.imtech.res.in/raghava/cbtope/.

DOI: 10.1186/1745-7580-6-6

PMCID: PMC2974664

PMID: 20961417

1859. Bioinformatics. 2010 Oct 15;26(20):2629-30. doi: 10.1093/bioinformatics/btq491.

Epub 2010 Aug 24.

Structural genomics of histone tail recognition.

Wang M(1), Mok MW, Harper H, Lee WH, Min J, Knapp S, Oppermann U, Marsden B,

Schapira M.

Author information:

(1)Structural Genomics Consortium, University of Oxford, Headington, Oxford

OX37DQ, UK.

SUMMARY: The structural genomics of histone tail recognition web server is an

open access resource that presents within mini articles all publicly available

experimental structures of histone tails in complex with human proteins. Each

article is composed of interactive 3D slides that dissect the structural

mechanism underlying the recognition of specific sequences and histone marks. A

concise text html-linked to interactive graphics guides the reader through the

main features of the interaction. This resource can be used to analyze and

compare binding modes across multiple histone recognition modules, to evaluate

the chemical tractability of binding sites involved in epigenetic signaling and

design small molecule inhibitors.

AVAILABILITY: http://www.thesgc.org/resources/histone\_tails/

CONTACT: matthieu.schapira@utoronto.ca

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq491

PMCID: PMC2951094

PMID: 20739309 [Indexed for MEDLINE]

1860. Bioinformatics. 2010 Oct 15;26(20):2624-5. doi: 10.1093/bioinformatics/btq480.

Epub 2010 Aug 24.

2Struc: the secondary structure server.

Klose DP(1), Wallace BA, Janes RW.

Author information:

(1)School of Biological and Chemical Sciences, Queen Mary, University of London,

Mile End Road, London E1 4NS, UK.

SUMMARY: The defined secondary structure of proteins method is often considered

the gold standard for assignment of secondary structure from three-dimensional

coordinates. However, there are alternative methods. '2Struc: The Secondary

Structure Server' has been created as a single point of access for eight

different secondary structure assignment methods. It has been designed to enable

comparisons between methods for analyzing the secondary structure content for a

single protein. It also includes a second functionality, 'Compare-the-Protein' to

enable comparisons of the secondary structure features from any one method to be

made within a collection of nuclear magnetic resonance models, or between the

crystal structures of two different proteins.

AVAILABILITY: http://2struc.cryst.bbk.ac.uk

CONTACT: r.w.janes@qmul.ac.uk

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq480

PMCID: PMC2951091

PMID: 20739308 [Indexed for MEDLINE]

1861. Bioinformatics. 2010 Oct 15;26(20):2649-50. doi: 10.1093/bioinformatics/btq487.

Epub 2010 Aug 24.

ICPS: an integrative cancer profiler system.

Zhang XY(1), Shi L, Liu Y, Tian F, Zhao HT, Miao XP, Huang ML, Zhu XY.

Author information:

(1)State Key Laboratory of Intelligent Technology and Systems, Tsinghua National

Laboratory for Information Science and Technology, Department of Computer Science

and Technology, Tsinghua University, Beijing 100084, China.

zhxy@mail.tsinghua.edu.cn

Founded upon the database of 570 public signatures, ICPS is a web-based

application to obtain biomarker profiles among 11 common cancers by integrating

genomic alterations with transcription signatures on the basis of a previously

developed integrative pipeline. ICPS supports both public data and user's

in-house data, and performs meta-analysis at a cancer subtype level by combining

heterogeneous datasets. Finally, ICPS returns the robust gene signature

containing potential cancer biomarkers that may be useful to carcinogenesis study

and clinical cancer diagnosis.AVAILABILITY: http://server.bioicps.org

CONTACT: zhxy@mail.tsinghua.edu.cn; zxy-dcs@mail.tsinghua.edu.cn

DOI: 10.1093/bioinformatics/btq487

PMID: 20736342 [Indexed for MEDLINE]

1862. BMC Bioinformatics. 2010 Oct 7;11 Suppl 6:S25. doi: 10.1186/1471-2105-11-S6-S25.

Integrated database for identifying candidate genes for Aspergillus flavus

resistance in maize.

Kelley RY(1), Gresham C, Harper J, Bridges SM, Warburton ML, Hawkins LK,

Pechanova O, Peethambaran B, Pechan T, Luthe DS, Mylroie JE, Ankala A, Ozkan S,

Henry WB, Williams WP.

Author information:

(1)Department of Biochemistry and Molecular Biology, Mississippi State

University, MS, USA. rkelley@bch.msstate.edu

BACKGROUND: Aspergillus flavus Link:Fr, an opportunistic fungus that produces

aflatoxin, is pathogenic to maize and other oilseed crops. Aflatoxin is a potent

carcinogen, and its presence markedly reduces the value of grain. Understanding

and enhancing host resistance to A. flavus infection and/or subsequent aflatoxin

accumulation is generally considered an efficient means of reducing grain losses

to aflatoxin. Different proteomic, genomic and genetic studies of maize (Zea mays

L.) have generated large data sets with the goal of identifying genes responsible

for conferring resistance to A. flavus, or aflatoxin.

RESULTS: In order to maximize the usage of different data sets in new studies,

including association mapping, we have constructed a relational database with web

interface integrating the results of gene expression, proteomic (both gel-based

and shotgun), Quantitative Trait Loci (QTL) genetic mapping studies, and sequence

data from the literature to facilitate selection of candidate genes for continued

investigation. The Corn Fungal Resistance Associated Sequences Database

(CFRAS-DB) (http://agbase.msstate.edu/) was created with the main goal of

identifying genes important to aflatoxin resistance. CFRAS-DB is implemented

using MySQL as the relational database management system running on a Linux

server, using an Apache web server, and Perl CGI scripts as the web interface.

The database and the associated web-based interface allow researchers to examine

many lines of evidence (e.g. microarray, proteomics, QTL studies, SNP data) to

assess the potential role of a gene or group of genes in the response of

different maize lines to A. flavus infection and subsequent production of

aflatoxin by the fungus.

CONCLUSIONS: CFRAS-DB provides the first opportunity to integrate data pertaining

to the problem of A. flavus and aflatoxin resistance in maize in one resource and

to support queries across different datasets. The web-based interface gives

researchers different query options for mining the database across different

types of experiments. The database is publically available at

http://agbase.msstate.edu.

DOI: 10.1186/1471-2105-11-S6-S25

PMCID: PMC3026372

PMID: 20946609 [Indexed for MEDLINE]

1863. Bioinformatics. 2010 Oct 1;26(19):2464-5. doi: 10.1093/bioinformatics/btq446.

Epub 2010 Aug 23.

GolgiP: prediction of Golgi-resident proteins in plants.

Chou WC(1), Yin Y, Xu Y.

Author information:

(1)Computational Systems Biology Laboratory, Department of Biochemistry and

Molecular Biology, and Institute of Bioinformatics, University of Georgia,

BioEnergy Science Center, GA, USA.

We present a novel Golgi-prediction server, GolgiP, for computational prediction

of both membrane- and non-membrane-associated Golgi-resident proteins in plants.

We have employed a support vector machine-based classification method for the

prediction of such Golgi proteins, based on three types of information, dipeptide

composition, transmembrane domain(s) (TMDs) and functional domain(s) of a

protein, where the functional domain information is generated through searching

against the Conserved Domains Database, and the TMD information includes the

number of TMDs, the length of TMD and the number of TMDs at the N-terminus of a

protein. Using GolgiP, we have made genome-scale predictions of Golgi-resident

proteins in 18 plant genomes, and have made the preliminary analysis of the

predicted data.AVAILABILITY: The GolgiP web service is publically available at

http://csbl1.bmb.uga.edu/GolgiP/.

DOI: 10.1093/bioinformatics/btq446

PMCID: PMC2944200

PMID: 20733061 [Indexed for MEDLINE]

1864. Bioinformatics. 2010 Oct 1;26(19):2474-6. doi: 10.1093/bioinformatics/btq452.

Epub 2010 Aug 10.

Genevar: a database and Java application for the analysis and visualization of

SNP-gene associations in eQTL studies.

Yang TP(1), Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE,

Deloukas P, Dermitzakis ET.

Author information:

(1)Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton,

Cambridge CB10 1HH, UK.

Genevar (GENe Expression VARiation) is a database and Java tool designed to

integrate multiple datasets, and provides analysis and visualization of

associations between sequence variation and gene expression. Genevar allows

researchers to investigate expression quantitative trait loci (eQTL) associations

within a gene locus of interest in real time. The database and application can be

installed on a standard computer in database mode and, in addition, on a server

to share discoveries among affiliations or the broader community over the

Internet via web services protocols.AVAILABILITY:

http://www.sanger.ac.uk/resources/software/genevar.

DOI: 10.1093/bioinformatics/btq452

PMCID: PMC2944204

PMID: 20702402 [Indexed for MEDLINE]

1865. Breast Cancer Res Treat. 2010 Oct;123(3):725-31. doi: 10.1007/s10549-009-0674-9.

Epub 2009 Dec 18.

An online survival analysis tool to rapidly assess the effect of 22,277 genes on

breast cancer prognosis using microarray data of 1,809 patients.

Györffy B(1), Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, Szallasi Z.

Author information:

(1)Joint Research Laboratory of the Hungarian Academy of Sciences and the

Semmelweis University, Semmelweis University 1st Department of Pediatrics, Bokay

u. 53-54, 1083, Budapest, Hungary. zsalab2@yahoo.com

Validating prognostic or predictive candidate genes in appropriately powered

breast cancer cohorts are of utmost interest. Our aim was to develop an online

tool to draw survival plots, which can be used to assess the relevance of the

expression levels of various genes on the clinical outcome both in untreated and

treated breast cancer patients. A background database was established using gene

expression data and survival information of 1,809 patients downloaded from GEO

(Affymetrix HGU133A and HGU133+2 microarrays). The median relapse free survival

is 6.43 years, 968/1,231 patients are estrogen-receptor (ER) positive, and

190/1,369 are lymph-node positive. After quality control and normalization only

probes present on both Affymetrix platforms were retained (n = 22,277). In order

to analyze the prognostic value of a particular gene, the cohorts are divided

into two groups according to the median (or upper/lower quartile) expression of

the gene. The two groups can be compared in terms of relapse free survival,

overall survival, and distant metastasis free survival. A survival curve is

displayed, and the hazard ratio with 95% confidence intervals and logrank P value

are calculated and displayed. Additionally, three subgroups of patients can be

assessed: systematically untreated patients, endocrine-treated ER positive

patients, and patients with a distribution of clinical characteristics

representative of those seen in general clinical practice in the US. Web address:

www.kmplot.com . We used this integrative data analysis tool to confirm the

prognostic power of the proliferation-related genes TOP2A and TOP2B, MKI67,

CCND2, CCND3, CCNDE2, as well as CDKN1A, and TK2. We also validated the

capability of microarrays to determine estrogen receptor status in 1,231

patients. The tool is highly valuable for the preliminary assessment of

biomarkers, especially for research groups with limited bioinformatic resources.

DOI: 10.1007/s10549-009-0674-9

PMID: 20020197 [Indexed for MEDLINE]

1866. J Biomol Struct Dyn. 2010 Oct;28(2):175-86.

Virus-mPLoc: a fusion classifier for viral protein subcellular location

prediction by incorporating multiple sites.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, 800 Dongchuan Road, Shanghai, 200240, China. hbshen@sjtu.edu.cn

Knowledge of the subcellular localization of viral proteins in a host cell or

virus-infected cell is very important because it is closely related to their

destructive tendencies and consequences. Facing the avalanche of new protein

sequences discovered in the post genomic era, we are challenged to develop

automated methods for quickly and accurately predicting the location sites of

viral proteins in a host cell; the information thus acquired is particularly

important for medical science and antiviral drug design. In view of this, a new

fusion classifier called "Virus-mPLoc" was established by hybridizing the gene

ontology information, functional domain information, and sequential evolutionary

information. The new predictor not only can more accurately predict the location

sites of viral proteins in a host cell, but also have the capacity to identify

the multiple-location virus proteins, which is beyond the reach of any existing

predictors specialized for viral proteins. For reader's convenience, a

user-friendly web-server for Virus-mPLoc was designed that is freely accessible

at http://www.csbio.sjtu.edu.cn/bioinf/virus-multi/.

DOI: 10.1080/07391102.2010.10507351

PMID: 20645651 [Indexed for MEDLINE]

1867. J Proteome Res. 2010 Oct 1;9(10):4992-5001. doi: 10.1021/pr100618t.

Large-scale prediction of human protein-protein interactions from amino acid

sequence based on latent topic features.

Pan XY(1), Zhang YN, Shen HB.

Author information:

(1)Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong

University, Shanghai, China.

Protein-protein interaction (PPI) is at the core of the entire interactomic

system of any living organism. Although there are many human protein-protein

interaction links being experimentally determined, the number is still relatively

very few compared to the estimation that there are ∼300,000 protein-protein

interactions in human beings. Hence, it is still urgent and challenging to

develop automated computational methods to accurately and efficiently predict

protein-protein interactions. In this paper, we propose a novel hierarchical

LDA-RF (latent dirichlet allocation-random forest) model to predict human

protein-protein interactions from protein primary sequences directly, which is

featured by a high success rate and strong ability for handling large-scale data

sets by digging the hidden internal structures buried into the noisy amino acid

sequences in low dimensional latent semantic space. First, the local sequential

features represented by conjoint triads are constructed from sequences. Then the

generative LDA model is used to project the original feature space into the

latent semantic space to obtain low dimensional latent topic features, which

reflect the hidden structures between proteins. Finally, the powerful random

forest model is used to predict the probability for interaction of two proteins.

Our results show that the proposed latent topic feature is very promising for PPI

prediction and could also become a powerful strategy to deal with many other

bioinformatics problems. As a web server, LDA-RF is freely available at

http://www.csbio.sjtu.edu.cn/bioinf/LR\_PPI for academic use.

DOI: 10.1021/pr100618t

PMID: 20698572 [Indexed for MEDLINE]

1868. Immunome Res. 2010 Sep 27;6 Suppl 1:S6. doi: 10.1186/1745-7580-6-S1-S6.

TAP Hunter: a SVM-based system for predicting TAP ligands using local description

of amino acid sequence.

Lam TH(1), Mamitsuka H, Ren EC, Tong JC.

Author information:

(1)Laboratory of Immunogenetics and Viral Host-Pathogen Genomics, Singapore

Immunology Network, 8A Biomedical Grove, #03-06, Immunos, Singapore 138648.

Lam\_Tze\_Hau@immunol.a-star.edu.sg.

BACKGROUND: Selective peptide transport by the transporter associated with

antigen processing (TAP) represents one of the main candidate mechanisms that may

regulate the presentation of antigenic peptides to HLA class I molecules. Because

TAP-binding preferences may significant impact T-cell epitope selection, there is

great interest in applying computational techniques to systematically discover

these elements.

RESULTS: We describe TAP Hunter, a web-based computational system for predicting

TAP-binding peptides. A novel encoding scheme, based on representations of TAP

peptide fragments and composition effects, allows the identification of

variable-length TAP ligands using SVM as the prediction engine. The system was

rigorously trained and tested using 613 experimentally verified peptide

sequences. The results showed that the system has good predictive ability with

area under the receiver operating characteristics curve (AROC) ≥0.88. In

addition, TAP Hunter is compared against several existing public available TAP

predictors and has showed either superior or comparable performance.

CONCLUSIONS: TAP Hunter provides a reliable platform for predicting variable

length peptides binding onto the TAP transporter. To facilitate the usage of TAP

Hunter to the scientific community, a simple, flexible and user-friendly

web-server is developed and freely available at

http://datam.i2r.a-star.edu.sg/taphunter/.

DOI: 10.1186/1745-7580-6-S1-S6

PMCID: PMC2946784

PMID: 20875157

1869. BMC Bioinformatics. 2010 Sep 22;11:473. doi: 10.1186/1471-2105-11-473.

PeptideMine--a webserver for the design of peptides for protein-peptide binding

studies derived from protein-protein interactomes.

Shameer K(1), Madan LL, Veeranna S, Gopal B, Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences (TIFR), GKVK Campus, Bellary Road,

Bangalore, 560065, India.

BACKGROUND: Signal transduction events often involve transient, yet specific,

interactions between structurally conserved protein domains and polypeptide

sequences in target proteins. The identification and validation of these

associating domains is crucial to understand signal transduction pathways that

modulate different cellular or developmental processes. Bioinformatics strategies

to extract and integrate information from diverse sources have been shown to

facilitate the experimental design to understand complex biological events. These

methods, primarily based on information from high-throughput experiments, have

also led to the identification of new connections thus providing hypothetical

models for cellular events. Such models, in turn, provide a framework for

directing experimental efforts for validating the predicted molecular rationale

for complex cellular processes. In this context, it is envisaged that the

rational design of peptides for protein-peptide binding studies could

substantially facilitate the experimental strategies to evaluate a predicted

interaction. This rational design procedure involves the integration of

protein-protein interaction data, gene ontology, physico-chemical calculations,

domain-domain interaction data and information on functional sites or critical

residues.

RESULTS: Here we describe an integrated approach called "PeptideMine" for the

identification of peptides based on specific functional patterns present in the

sequence of an interacting protein. This approach based on sequence searches in

the interacting sequence space has been developed into a webserver, which can be

used for the identification and analysis of peptides, peptide homologues or

functional patterns from the interacting sequence space of a protein. To further

facilitate experimental validation, the PeptideMine webserver also provides a

list of physico-chemical parameters corresponding to the peptide to determine the

feasibility of using the peptide for in vitro biochemical or biophysical studies.

CONCLUSIONS: The strategy described here involves the integration of data and

tools to identify potential interacting partners for a protein and design

criteria for peptides based on desired biochemical properties. Alongside the

search for interacting protein sequences using three different search programs,

the server also provides the biochemical characteristics of candidate peptides to

prune peptide sequences based on features that are most suited for a given

experiment. The PeptideMine server is available at the URL:

http://caps.ncbs.res.in/peptidemine.

DOI: 10.1186/1471-2105-11-473

PMCID: PMC2955050

PMID: 20858292 [Indexed for MEDLINE]

1870. BMC Genomics. 2010 Sep 22;11:507. doi: 10.1186/1471-2164-11-507.

Prediction and classification of aminoacyl tRNA synthetases using PROSITE

domains.

Panwar B(1), Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India.

BACKGROUND: Aminoacyl tRNA synthetases (aaRSs) catalyse the first step of protein

synthesis in all organisms. They are responsible for the precise attachment of

amino acids to their cognate transfer RNAs. There are twenty different types of

aaRSs, unique for each amino acid. These aaRSs have been divided into two

classes, each comprising ten enzymes. It is important to predict and classify

aaRSs in order to understand protein synthesis.

RESULTS: In this study, all models were developed on a non-redundant dataset

containing 117 aaRSs and an equal number of non-aaRSs, in which no two sequences

have more than 30% similarity. First, we applied the similarity search technique,

BLAST, and achieved a maximum accuracy of 67.52%. We observed that 62% of tRNA

synthetases contain one or more domains from amongst the following four PROSITE

domains: PS50862, PS00178, PS50860 and PS50861. An SVM-based model was developed

to discriminate between aaRSs, and non-aaRSs, and achieved a maximum MCC of 0.68

with accuracy of 83.73%, using selective dipeptide composition. We developed a

hybrid approach and achieved a maximum MCC of 0.72 with accuracy of 85.49%, where

SVM model developed using selected dipeptide composition and information of four

PROSITE domains. We further developed an SVM-based model for classifying the

aaRSs into class-1 and class-2, using selective dipeptide composition and

achieved an MCC of 0.79. We also observed that two domains (PS00178, PS50889) in

class-1 and three domains (PS50862, PS50860, PS50861) in class-2 were preferred.

A hybrid method was developed using these domains as descriptor, along with

selected dipeptide composition, and achieved an MCC of 0.87 with a sensitivity of

94.55% and an accuracy of 93.19%. All models were evaluated using a five-fold

cross-validation technique.

CONCLUSIONS: We have analyzed protein sequences of aaRSs (class-1 and class-2)

and non-aaRSs and identified interesting patterns. The high accuracy achieved by

our SVM models using selected dipeptide composition demonstrates that certain

types of dipeptide are preferred in aaRSs. We were able to identify PROSITE

domains that are preferred in aaRSs and their classes, providing interesting

insights into tRNA synthetases. The method developed in this study will be useful

for researchers studying aaRS enzymes and tRNA biology. The web-server based on

the above study, is available at http://www.imtech.res.in/raghava/icaars/.

DOI: 10.1186/1471-2164-11-507

PMCID: PMC2997003

PMID: 20860794 [Indexed for MEDLINE]

1871. Bioinformatics. 2010 Sep 15;26(18):2338-9. doi: 10.1093/bioinformatics/btq337.

Epub 2010 Jun 23.

Chromaligner: a web server for chromatogram alignment.

Wang SY(1), Ho TJ, Kuo CH, Tseng YJ.

Author information:

(1)Department of Computer Science and Information Engineering, College of

Medicine, National Taiwan University, Taipei, Taiwan.

Chromaligner is a tool for chromatogram alignment to align retention time for

chromatographic methods coupled to spectrophotometers such as high performance

liquid chromatography and capillary electrophoresis for metabolomics works.

Chromaligner resolves peak shifts by a constrained chromatogram alignment. For a

collection of chromatograms and a set of defined peaks, Chromaligner aligns the

chromatograms on defined peaks using correlation warping (COW). Chromaligner is

faster than the original COW algorithm by k(2) times, where k is the number of

defined peaks in a chromatogram. It also provides alignments based on known

component peaks to reach the best results for further chemometric

analysis.AVAILABILITY: Chromaligner is freely accessible at

http://cmdd.csie.ntu.edu.tw/~chromaligner.

DOI: 10.1093/bioinformatics/btq337

PMID: 20576623 [Indexed for MEDLINE]

1872. BMC Bioinformatics. 2010 Sep 15;11:461. doi: 10.1186/1471-2105-11-461.

Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis

of core and accessory genomic regions.

Laing C(1), Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, Thomas JE,

Gannon VP.

Author information:

(1)Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, Lethbridge,

AB, Canada.

BACKGROUND: The pan-genome of a bacterial species consists of a core and an

accessory gene pool. The accessory genome is thought to be an important source of

genetic variability in bacterial populations and is gained through lateral gene

transfer, allowing subpopulations of bacteria to better adapt to specific niches.

Low-cost and high-throughput sequencing platforms have created an exponential

increase in genome sequence data and an opportunity to study the pan-genomes of

many bacterial species. In this study, we describe a new online pan-genome

sequence analysis program, Panseq.

RESULTS: Panseq was used to identify Escherichia coli O157:H7 and E. coli K-12

genomic islands. Within a population of 60 E. coli O157:H7 strains, the existence

of 65 accessory genomic regions identified by Panseq analysis was confirmed by

PCR. The accessory genome and binary presence/absence data, and core genome and

single nucleotide polymorphisms (SNPs) of six L. monocytogenes strains were

extracted with Panseq and hierarchically clustered and visualized. The nucleotide

core and binary accessory data were also used to construct maximum parsimony (MP)

trees, which were compared to the MP tree generated by multi-locus sequence

typing (MLST). The topology of the accessory and core trees was identical but

differed from the tree produced using seven MLST loci. The Loci Selector module

found the most variable and discriminatory combinations of four loci within a 100

loci set among 10 strains in 1 s, compared to the 449 s required to exhaustively

search for all possible combinations; it also found the most discriminatory 20

loci from a 96 loci E. coli O157:H7 SNP dataset.

CONCLUSION: Panseq determines the core and accessory regions among a collection

of genomic sequences based on user-defined parameters. It readily extracts

regions unique to a genome or group of genomes, identifies SNPs within shared

core genomic regions, constructs files for use in phylogeny programs based on

both the presence/absence of accessory regions and SNPs within core regions and

produces a graphical overview of the output. Panseq also includes a loci selector

that calculates the most variable and discriminatory loci among sets of accessory

loci or core gene SNPs.

AVAILABILITY: Panseq is freely available online at http://76.70.11.198/panseq.

Panseq is written in Perl.

DOI: 10.1186/1471-2105-11-461

PMCID: PMC2949892

PMID: 20843356 [Indexed for MEDLINE]

1873. BMC Bioinformatics. 2010 Sep 9;11:455. doi: 10.1186/1471-2105-11-455.

Next generation tools for genomic data generation, distribution, and

visualization.

Nix DA(1), Di Sera TL, Dalley BK, Milash BA, Cundick RM, Quinn KS, Courdy SJ.

Author information:

(1)Department of Oncological Sciences, University of Utah, Huntsman Cancer

Institute, Salt Lake City, USA. david.nix@hci.utah.edu

BACKGROUND: With the rapidly falling cost and availability of high throughput

sequencing and microarray technologies, the bottleneck for effectively using

genomic analysis in the laboratory and clinic is shifting to one of effectively

managing, analyzing, and sharing genomic data.

RESULTS: Here we present three open-source, platform independent, software tools

for generating, analyzing, distributing, and visualizing genomic data. These

include a next generation sequencing/microarray LIMS and analysis project center

(GNomEx); an application for annotating and programmatically distributing genomic

data using the community vetted DAS/2 data exchange protocol (GenoPub); and a

standalone Java Swing application (GWrap) that makes cutting edge command line

analysis tools available to those who prefer graphical user interfaces. Both

GNomEx and GenoPub use the rich client Flex/Flash web browser interface to

interact with Java classes and a relational database on a remote server. Both

employ a public-private user-group security model enabling controlled

distribution of patient and unpublished data alongside public resources. As such,

they function as genomic data repositories that can be accessed manually or

programmatically through DAS/2-enabled client applications such as the Integrated

Genome Browser.

CONCLUSIONS: These tools have gained wide use in our core facilities, research

laboratories and clinics and are freely available for non-profit use. See

http://sourceforge.net/projects/gnomex/,

http://sourceforge.net/projects/genoviz/, and

http://sourceforge.net/projects/useq.

DOI: 10.1186/1471-2105-11-455

PMCID: PMC2944281

PMID: 20828407 [Indexed for MEDLINE]

1874. Bioinformatics. 2010 Sep 1;26(17):2153-9. doi: 10.1093/bioinformatics/btq341.

Epub 2010 Jul 22.

Prophossi: automating expert validation of phosphopeptide-spectrum matches from

tandem mass spectrometry.

Martin DM(1), Nett IR, Vandermoere F, Barber JD, Morrice NA, Ferguson MA.

Author information:

(1)Division of Biological Chemistry and Drug Discovery, College of LifeSciences,

University of Dundee, Dow Street, Dundee DD1 5EH, UK. d.m.a.martin@dundee.ac.uk

MOTIVATION: Complex patterns of protein phosphorylation mediate many cellular

processes. Tandem mass spectrometry (MS/MS) is a powerful tool for identifying

these post-translational modifications. In high-throughput experiments, mass

spectrometry database search engines, such as MASCOT provide a ranked list of

peptide identifications based on hundreds of thousands of MS/MS spectra obtained

in a mass spectrometry experiment. These search results are not in themselves

sufficient for confident assignment of phosphorylation sites as identification of

characteristic mass differences requires time-consuming manual assessment of the

spectra by an experienced analyst. The time required for manual assessment has

previously rendered high-throughput confident assignment of phosphorylation sites

challenging.

RESULTS: We have developed a knowledge base of criteria, which replicate expert

assessment, allowing more than half of cases to be automatically validated and

site assignments verified with a high degree of confidence. This was assessed by

comparing automated spectral interpretation with careful manual examination of

the assignments for 501 peptides above the 1% false discovery rate (FDR)

threshold corresponding to 259 putative phosphorylation sites in 74 proteins of

the Trypanosoma brucei proteome. Despite this stringent approach, we are able to

validate 80 of the 91 phosphorylation sites (88%) positively identified by manual

examination of the spectra used for the MASCOT searches with a FDR < 15%.

CONCLUSIONS: High-throughput computational analysis can provide a viable second

stage validation of primary mass spectrometry database search results. Such

validation gives rapid access to a systems level overview of protein

phosphorylation in the experiment under investigation.

AVAILABILITY: A GPL licensed software implementation in Perl for analysis and

spectrum annotation is available in the supplementary material and a web server

can be assessed online at http://www.compbio.dundee.ac.uk/prophossi.

DOI: 10.1093/bioinformatics/btq341

PMCID: PMC2922888

PMID: 20651112 [Indexed for MEDLINE]

1875. Gene. 2010 Sep 1;463(1-2):1-7. doi: 10.1016/j.gene.2010.04.012. Epub 2010 May 2.

Conserved miRNAs and their targets identified in lettuce (Lactuca) by EST

analysis.

Han Y(1), Zhu B, Luan F, Zhu H, Shao Y, Chen A, Lu C, Luo Y.

Author information:

(1)Laboratory of Fruit Biology, College of Food Science and Nutritional

Engineering, China Agricultural University, Beijing 100083, China.

MicroRNAs (miRNAs) are a newly identified class of endogenous, non-coding, short

( approximately 21nt) RNAs that play important roles in regulating gene

expression at post-transcriptional level by targeting mRNA cleavage or

translational inhibition in plants and animals. Though there are lots of

differences between plant miRNAs and animal miRNAs, most of these tiny RNAs are

highly conserved in each kingdom. Here, we show the conserved miRNAs in lettuce

(Lactuca) identified using EST (expressed sequence tag) analysis. Namely, all

previously known miRNAs in other plant species were blasted against lettuce EST

sequences to select novel miRNAs in lettuce by a series of filtering criteria. By

this strategy, we found a total of 21 conserved miRNAs belonging to 12 miRNA

families. After analyzing the conservation and evolution of lettuce miRNAs and

their counterparts in other plant species, we revealed that though miRNAs are

highly conserved, some specific sites are more likely to mutate. To confirm the

expression of identified miRNAs in lettuce, an RT-PCR approach was employed.

Moreover, all identified lettuce miRNAs were used to search their potential

target genes by miRU web-server from TIGR database available at

http://www.tigr.org and a total of 63 potential targets for 10 identified miRNA

families in lettuce were found. Similar to previous works, some miRNA targets are

transcription factors involved in lettuce growth and development, metabolism, and

stress responses.

Copyright 2010 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.gene.2010.04.012

PMID: 20441788 [Indexed for MEDLINE]

1876. J Clin Microbiol. 2010 Sep;48(9):3388-91. doi: 10.1128/JCM.00921-10. Epub 2010

Jul 21.

Ability of Bacillus cereus group strains to cause food poisoning varies according

to phylogenetic affiliation (groups I to VII) rather than species affiliation.

Guinebretière MH(1), Velge P, Couvert O, Carlin F, Debuyser ML, Nguyen-The C.

Author information:

(1)INRA, UMR A408, Université d'Avignon, Domaine Saint-Paul, Site Agroparc,

F-84914 Avignon Cedex 9, France. guinebre@avignon.inra.fr

Cytotoxic activity levels of culture filtrates and toxin distributions varied

according to the phylogenetic group (I to VII) within the Bacillus cereus group,

suggesting that these groups are of different clinical significance and are more

suitable than species affiliations for determining food poisoning risk. A

first-line, simple online tool

(https://www.tools.symprevius.org/Bcereus/english.php) to assign strains to the

different phylogenetic groups is presented.

DOI: 10.1128/JCM.00921-10

PMCID: PMC2937725

PMID: 20660215 [Indexed for MEDLINE]

1877. Nurse Educ. 2010 Sep-Oct;35(5):205-7. doi: 10.1097/NNE.0b013e3181ed81e4.

Zotero: harnessing the power of a personal bibliographic manager.

Coar JT(1), Sewell JP.

Author information:

(1)School of Nursing, Georgia College & State University, Milledgeville, Georgia

61061, USA.

Zotero is a powerful free personal bibliographic manager (PBM) for writers. Use

of a PBM allows the writer to focus on content, rather than the tedious details

of formatting citations and references. Zotero 2.0 (http://www.zotero.org) has

new features including the ability to synchronize citations with the off-site

Zotero server and the ability to collaborate and share with others. An overview

on how to use the software and discussion about the strengths and limitations are

included.

DOI: 10.1097/NNE.0b013e3181ed81e4

PMID: 20729678 [Indexed for MEDLINE]

1878. BMC Bioinformatics. 2010 Aug 27;11:439. doi: 10.1186/1471-2105-11-439.

CMASA: an accurate algorithm for detecting local protein structural similarity

and its application to enzyme catalytic site annotation.

Li GH(1), Huang JF.

Author information:

(1)State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of

Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China.

BACKGROUND: The rapid development of structural genomics has resulted in many

"unknown function" proteins being deposited in Protein Data Bank (PDB), thus, the

functional prediction of these proteins has become a challenge for structural

bioinformatics. Several sequence-based and structure-based methods have been

developed to predict protein function, but these methods need to be improved

further, such as, enhancing the accuracy, sensitivity, and the computational

speed. Here, an accurate algorithm, the CMASA (Contact MAtrix based local

Structural Alignment algorithm), has been developed to predict unknown functions

of proteins based on the local protein structural similarity. This algorithm has

been evaluated by building a test set including 164 enzyme families, and also

been compared to other methods.

RESULTS: The evaluation of CMASA shows that the CMASA is highly accurate (0.96),

sensitive (0.86), and fast enough to be used in the large-scale functional

annotation. Comparing to both sequence-based and global structure-based methods,

not only the CMASA can find remote homologous proteins, but also can find the

active site convergence. Comparing to other local structure comparison-based

methods, the CMASA can obtain the better performance than both FFF (a method

using geometry to predict protein function) and SPASM (a local structure

alignment method); and the CMASA is more sensitive than PINTS and is more

accurate than JESS (both are local structure alignment methods). The CMASA was

applied to annotate the enzyme catalytic sites of the non-redundant PDB, and at

least 166 putative catalytic sites have been suggested, these sites can not be

observed by the Catalytic Site Atlas (CSA).

CONCLUSIONS: The CMASA is an accurate algorithm for detecting local protein

structural similarity, and it holds several advantages in predicting enzyme

active sites. The CMASA can be used in large-scale enzyme active site annotation.

The CMASA can be available by the mail-based server

(http://159.226.149.45/other1/CMASA/CMASA.htm).

DOI: 10.1186/1471-2105-11-439

PMCID: PMC2936402

PMID: 20796320 [Indexed for MEDLINE]

1879. BMC Bioinformatics. 2010 Aug 23;11:435. doi: 10.1186/1471-2105-11-435.

CircuitsDB: a database of mixed microRNA/transcription factor feed-forward

regulatory circuits in human and mouse.

Friard O(1), Re A, Taverna D, De Bortoli M, Corá D.

Author information:

(1)Center for Molecular Systems Biology, University of Torino, Via Accademia

Albertina, 13 - I-10123 Torino, Italy.

BACKGROUND: Transcription Factors (TFs) and microRNAs (miRNAs) are key players

for gene expression regulation in higher eukaryotes. In the last years, a large

amount of bioinformatic studies were devoted to the elucidation of

transcriptional and post-transcriptional (mostly miRNA-mediated) regulatory

interactions, but little is known about the interplay between them.

DESCRIPTION: Here we describe a dynamic web-accessible database, CircuitsDB,

supporting a genome-wide transcriptional and post-transcriptional regulatory

network integration, for the human and mouse genomes, based on a bioinformatic

sequence-analysis approach. In particular, CircuitsDB is currently focused on the

study of mixed miRNA/TF Feed-Forward regulatory Loops (FFLs), i.e. elementary

circuits in which a master TF regulates an miRNA and together with it a set of

Joint Target protein-coding genes. The database was constructed using an

ab-initio oligo analysis procedure for the identification of the transcriptional

and post-transcriptional interactions. Several external sources of information

were then pooled together to obtain the functional annotation of the proposed

interactions. Results for human and mouse genomes are presented in an integrated

web tool, that allows users to explore the circuits, investigate their sequence

and functional properties and thus suggest possible biological experiments.

CONCLUSIONS: We present CircuitsDB, a web-server devoted to the study of human

and mouse mixed miRNA/TF Feed-Forward regulatory circuits, freely available at:

http://biocluster.di.unito.it/circuits/

DOI: 10.1186/1471-2105-11-435

PMCID: PMC2936401

PMID: 20731828 [Indexed for MEDLINE]

1880. PLoS One. 2010 Aug 23;5(8):e12352. doi: 10.1371/journal.pone.0012352.

Presaging critical residues in protein interfaces-web server (PCRPi-W): a web

server to chart hot spots in protein interfaces.

Segura Mora J(1), Assi SA, Fernandez-Fuentes N.

Author information:

(1)Section of Experimental Therapeutics, Leeds Institute of Molecular Medicine,

University of Leeds, Leeds, United Kingdom.

BACKGROUND: It is well established that only a portion of residues that mediate

protein-protein interactions (PPIs), the so-called hot spot, contributes the most

to the total binding energy, and thus its identification is an important and

relevant question that has clear applications in drug discovery and protein

design. The experimental identification of hot spots is however a lengthy and

costly process, and thus there is an interest in computational tools that can

complement and guide experimental efforts.

PRINCIPAL FINDINGS: Here, we present Presaging Critical Residues in Protein

interfaces-Web server (http://www.bioinsilico.org/PCRPi), a web server that

implements a recently described and highly accurate computational tool designed

to predict critical residues in protein interfaces: PCRPi. PRCPi depends on the

integration of structural, energetic, and evolutionary-based measures by using

Bayesian Networks (BNs).

CONCLUSIONS: PCRPi-W has been designed to provide an easy and convenient access

to the broad scientific community. Predictions are readily available for download

or presented in a web page that includes among other information links to

relevant files, sequence information, and a Jmol applet to visualize and analyze

the predictions in the context of the protein structure.

DOI: 10.1371/journal.pone.0012352

PMCID: PMC2925954

PMID: 20808810 [Indexed for MEDLINE]

1881. BMC Bioinformatics. 2010 Aug 21;11:434. doi: 10.1186/1471-2105-11-434.

A classification approach for genotyping viral sequences based on

multidimensional scaling and linear discriminant analysis.

Kim J(1), Ahn Y, Lee K, Park SH, Kim S.

Author information:

(1)Department of Bioinformatics & Life Sciences, Soongsil University, Seoul,

Korea.

BACKGROUND: Accurate classification into genotypes is critical in understanding

evolution of divergent viruses. Here we report a new approach, MuLDAS, which

classifies a query sequence based on the statistical genotype models learned from

the known sequences. Thus, MuLDAS utilizes full spectra of well characterized

sequences as references, typically of an order of hundreds, in order to estimate

the significance of each genotype assignment.

RESULTS: MuLDAS starts by aligning the query sequence to the reference multiple

sequence alignment and calculating the subsequent distance matrix among the

sequences. They are then mapped to a principal coordinate space by

multidimensional scaling, and the coordinates of the reference sequences are used

as features in developing linear discriminant models that partition the space by

genotype. The genotype of the query is then given as the maximum a posteriori

estimate. MuLDAS tests the model confidence by leave-one-out cross-validation and

also provides some heuristics for the detection of 'outlier' sequences that fall

far outside or in-between genotype clusters. We have tested our method by

classifying HIV-1 and HCV nucleotide sequences downloaded from NCBI GenBank,

achieving the overall concordance rates of 99.3% and 96.6%, respectively, with

the benchmark test dataset retrieved from the respective databases of Los Alamos

National Laboratory.

CONCLUSIONS: The highly accurate genotype assignment coupled with several

measures for evaluating the results makes MuLDAS useful in analyzing the

sequences of rapidly evolving viruses such as HIV-1 and HCV. A web-based genotype

prediction server is available at http://www.muldas.org/MuLDAS/.

DOI: 10.1186/1471-2105-11-434

PMCID: PMC2936400

PMID: 20727194 [Indexed for MEDLINE]

1882. Bioinformatics. 2010 Aug 15;26(16):2066-8. doi: 10.1093/bioinformatics/btq324.

Epub 2010 Jun 23.

EpiTOP--a proteochemometric tool for MHC class II binding prediction.

Dimitrov I(1), Garnev P, Flower DR, Doytchinova I.

Author information:

(1)Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria.

MOTIVATION: T-cell epitope identification is a critical immunoinformatic problem

within vaccine design. To be an epitope, a peptide must bind an MHC protein.

RESULTS: Here, we present EpiTOP, the first server predicting MHC class II

binding based on proteochemometrics, a QSAR approach for ligands binding to

several related proteins. EpiTOP uses a quantitative matrix to predict binding to

12 HLA-DRB1 alleles. It identifies 89% of known epitopes within the top 20% of

predicted binders, reducing laboratory labour, materials and time by 80%. EpiTOP

is easy to use, gives comprehensive quantitative predictions and will be expanded

and updated with new quantitative matrices over time.

AVAILABILITY: EpiTOP is freely accessible at http://www.pharmfac.net/EpiTOP.

DOI: 10.1093/bioinformatics/btq324

PMID: 20576624 [Indexed for MEDLINE]

1883. J Pathol Inform. 2010 Aug 10;1. pii: 12. doi: 10.4103/2153-3539.68314.

A decade of experience in the development and implementation of tissue banking

informatics tools for intra and inter-institutional translational research.

Amin W(1), Singh H, Pople AK, Winters S, Dhir R, Parwani AV, Becich MJ.

Author information:

(1)Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh,

USA.

CONTEXT: Tissue banking informatics deals with standardized annotation,

collection and storage of biospecimens that can further be shared by researchers.

Over the last decade, the Department of Biomedical Informatics (DBMI) at the

University of Pittsburgh has developed various tissue banking informatics tools

to expedite translational medicine research. In this review, we describe the

technical approach and capabilities of these models.

DESIGN: Clinical annotation of biospecimens requires data retrieval from various

clinical information systems and the de-identification of the data by an honest

broker. Based upon these requirements, DBMI, with its collaborators, has

developed both Oracle-based organ-specific data marts and a more generic,

model-driven architecture for biorepositories. The organ-specific models are

developed utilizing Oracle 9.2.0.1 server tools and software applications and the

model-driven architecture is implemented in a J2EE framework.

RESULT: The organ-specific biorepositories implemented by DBMI include the

Cooperative Prostate Cancer Tissue Resource (http://www.cpctr.info/),

Pennsylvania Cancer Alliance Bioinformatics Consortium

(http://pcabc.upmc.edu/main.cfm), EDRN Colorectal and Pancreatic Neoplasm

Database (http://edrn.nci.nih.gov/) and Specialized Programs of Research

Excellence (SPORE) Head and Neck Neoplasm Database

(http://spores.nci.nih.gov/current/hn/index.htm). The model-based architecture is

represented by the National Mesothelioma Virtual Bank (http://mesotissue.org/).

These biorepositories provide thousands of well annotated biospecimens for the

researchers that are searchable through query interfaces available via the

Internet.

CONCLUSION: These systems, developed and supported by our institute, serve to

form a common platform for cancer research to accelerate progress in clinical and

translational research. In addition, they provide a tangible infrastructure and

resource for exposing research resources and biospecimen services in

collaboration with the clinical anatomic pathology laboratory information system

(APLIS) and the cancer registry information systems.

DOI: 10.4103/2153-3539.68314

PMCID: PMC2941965

PMID: 20922029

1884. J Proteomics. 2010 Aug 5;73(9):1740-6. doi: 10.1016/j.jprot.2010.05.011. Epub

2010 May 31.

Implementation and evaluation of relative and absolute quantification in shotgun

proteomics with label-free methods.

Grossmann J(1), Roschitzki B, Panse C, Fortes C, Barkow-Oesterreicher S,

Rutishauser D, Schlapbach R.

Author information:

(1)UZH ETH Zurich, Functional Genomics Center Zurich, Winterthurerstrasse 190,

CH-8057 Zurich, Switzerland. jg@fgcz.ethz.ch

Tandem mass spectrometry allows for fast protein identification in a complex

sample. As mass spectrometers get faster, more sensitive and more accurate,

methods were devised by many academic research groups and commercial suppliers

that allow protein research also in quantitative respect. Since label-free

methods are an attractive alternative to labeling approaches for proteomics

researchers seeking for accurate quantitative results we evaluated several

open-source analysis tools in terms of performance on two reference data sets,

explicitly generated for this purpose. In this paper we present an

implementation, T3PQ (Top 3 Protein Quantification), of the method suggested by

Silva and colleagues for LC-MS(E) applications and we demonstrate its

applicability to data generated on FT-ICR instruments acquiring in data dependent

acquisition (DDA) mode. In order to validate this method and to show its

usefulness also for absolute protein quantification, we generated a reference

data set of a sample containing four different proteins with known

concentrations. Furthermore, we compare three other label-free quantification

methods using a complex biological sample spiked with a standard protein in

defined concentrations. We evaluate the applicability of these methods and the

quality of the results in terms of robustness and dynamic range of the spiked-in

protein as well as other proteins also detected in the mixture. We discuss

drawbacks of each method individually and consider crucial points for

experimental designs. The source code of our implementation is available under

the terms of the GNU GPLv3 and can be downloaded from sourceforge

(http://fqms.svn.sourceforge.net/svnroot/fqms). A tarball containing the data

used for the evaluation is available on the FGCZ web server

(http://fgcz-data.uzh.ch/public/T3PQ.tgz).

(c) 2010 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.jprot.2010.05.011

PMID: 20576481 [Indexed for MEDLINE]

1885. Bioinformatics. 2010 Aug 1;26(15):1903-4. doi: 10.1093/bioinformatics/btq304.

Epub 2010 Jul 6.

Cloud-Coffee: implementation of a parallel consistency-based multiple alignment

algorithm in the T-Coffee package and its benchmarking on the Amazon

Elastic-Cloud.

Di Tommaso P(1), Orobitg M, Guirado F, Cores F, Espinosa T, Notredame C.

Author information:

(1)Centre For Genomic Regulation (Pompeu Fabra University), Carrer del Doctor

Aiguader 88, Barcelona, Spain.

SUMMARY: We present the first parallel implementation of the T-Coffee

consistency-based multiple aligner. We benchmark it on the Amazon Elastic Cloud

(EC2) and show that the parallelization procedure is reasonably effective. We

also conclude that for a web server with moderate usage (10K hits/month) the

cloud provides a cost-effective alternative to in-house deployment.

AVAILABILITY: T-Coffee is a freeware open source package available from

http://www.tcoffee.org/homepage.html

DOI: 10.1093/bioinformatics/btq304

PMCID: PMC2905555

PMID: 20605929 [Indexed for MEDLINE]

1886. Bioinformatics. 2010 Aug 1;26(15):1913-4. doi: 10.1093/bioinformatics/btq288.

Epub 2010 Jun 10.

PESDserv: a server for high-throughput comparison of protein binding site

surfaces.

Das S(1), Krein MP, Breneman CM.

Author information:

(1)Department of Chemistry and Chemical Biology, Rensselaer Polytechnic

Institute, 110 Eighth Street, Troy, NY 12180, USA.

SUMMARY: Structure-based approaches complement ligand-based approaches for

lead-discovery and cross-reactivity prediction. We present to the scientific

community a web server for comparing the surface of a ligand bound site of a

protein against a ligand bound site surface database of 106 796 sites. The web

server implements the property encoded shape distributions (PESD) algorithm for

surface comparison. A typical virtual screen takes 5 min to complete. The output

provides a ranked list of sites (by site similarity), hyperlinked to the

corresponding entries in the PDB and PDBeChem databases.

AVAILABILITY: The server is freely accessible at

http://reccr.chem.rpi.edu/Software/pesdserv/

DOI: 10.1093/bioinformatics/btq288

PMCID: PMC2905548

PMID: 20538727 [Indexed for MEDLINE]

1887. Bioinformatics. 2010 Aug 1;26(15):1920-1. doi: 10.1093/bioinformatics/btq298.

Epub 2010 Jun 6.

MiRror: a combinatorial analysis web tool for ensembles of microRNAs and their

targets.

Friedman Y(1), Naamati G, Linial M.

Author information:

(1)Department of Biological Chemistry, Sudarsky Center for Computational Biology

and School of Computer Science and Engineering, Hebrew University of Jerusalem,

Israel.

SUMMARY: The miRror application provides insights on microRNA (miRNA) regulation.

It is based on the notion of a combinatorial regulation by an ensemble of miRNAs

or genes. miRror integrates predictions from a dozen of miRNA resources that are

based on complementary algorithms into a unified statistical framework. For

miRNAs set as input, the online tool provides a ranked list of targets, based on

set of resources selected by the user, according to their significance of being

coordinately regulated. Symmetrically, a set of genes can be used as input to

suggest a set of miRNAs. The user can restrict the analysis for the preferred

tissue or cell line. miRror is suitable for analyzing results from miRNAs

profiling, proteomics and gene expression arrays.

AVAILABILITY: http://www.proto.cs.huji.ac.il/mirror

DOI: 10.1093/bioinformatics/btq298

PMID: 20529892 [Indexed for MEDLINE]

1888. Bioinformatics. 2010 Aug 1;26(15):1905-6. doi: 10.1093/bioinformatics/btq306.

Epub 2010 Jun 6.

COMA server for protein distant homology search.

Margelevicius M(1), Laganeckas M, Venclovas C.

Author information:

(1)Institute of Biotechnology, Graiciūno 8, Vilnius, Lithuania.

SUMMARY: Detection of distant homology is a widely used computational approach

for studying protein evolution, structure and function. Here, we report a

homology search web server based on sequence profile-profile comparison. The user

may perform searches in one of several regularly updated profile databases using

either a single sequence or a multiple sequence alignment as an input. The same

profile databases can also be downloaded for local use. The capabilities of the

server are illustrated with the identification of new members of the highly

diverse PD-(D/E)XK nuclease superfamily.

AVAILABILITY: http://www.ibt.lt/bioinformatics/coma/

DOI: 10.1093/bioinformatics/btq306

PMID: 20529888 [Indexed for MEDLINE]

1889. Heart. 2010 Aug;96(16):1264-7. doi: 10.1136/hrt.2009.192328.

The Myocardial Ischaemia National Audit Project (MINAP).

Herrett E(1), Smeeth L, Walker L, Weston C; MINAP Academic Group.

Collaborators: Birkhead J, Boyle R, Cunningham D, Eldridge S, Fox KA, Gale C,

Hemingway H, Squire I, Timmis A, Wilkinson P.

Author information:

(1)Department of Epidemiology and Population Health, Non-communicable Disease

Epidemiology Unit, London School of Hygiene and Tropical Medicine, London, UK.

emily.herrett@lshtm.ac.uk

AIMS OF MINAP: To audit the quality of care of patients with acute coronary

syndrome and provide a resource for academic research.QUALITY OF CARE

INTERVENTIONS: Feedback to hospitals, ambulance services and cardiac networks

regarding benchmarking of performance against national standards and targets.

SETTING: All 230 acute hospitals in England and Wales. Years: 2000-present.

POPULATION: Consecutive patients, unconsented. Current number of records: 735

000.

STARTPOINTS: Any acute coronary syndrome, including non-ST-elevation myocardial

infarction, ST-elevation myocardial infarction and unstable angina.

BASELINE DATA: 123 fields covering demographic factors, co-morbid conditions and

treatment in hospital. No blood resource.

DATA CAPTURE: Manual entry by clerks, nurses or doctors onto Lotus Notes;

non-financial incentives at hospital level.

DATA QUALITY: Hospitals perform an annual data validation study, where data are

re-entered from the case notes in 20 randomly selected records that are held on

the server. In 2008 data were >90% complete for 20 key fields, with >80%

completeness for all but four of the remaining fields.

ENDPOINTS AND LINKAGES TO OTHER DATA: All-cause mortality is obtained through

linkage with Office for National Statistics. No other linkages exist at present.

ACCESS TO DATA: Available for research and audit by application to the MINAP

Academic Group.

http://www.rcplondon.ac.uk/CLINICAL-STANDARDS/ORGANISATION/PARTNERSHIP/Pages/MINA

P-.aspx.

DOI: 10.1136/hrt.2009.192328

PMCID: PMC3505836

PMID: 20659944 [Indexed for MEDLINE]

1890. J Digit Imaging. 2010 Aug;23(4):454-62. doi: 10.1007/s10278-009-9200-1. Epub 2009

May 5.

Linking whole-slide microscope images with DICOM by using JPEG2000 interactive

protocol.

Tuominen VJ(1), Isola J.

Author information:

(1)Institute of Medical Technology, University of Tampere and Tampere University

Hospital, 33014, Tampere, Finland. vilppu.tuominen@uta.fi

The use of digitized histopathologic specimens (also known as whole-slide images

(WSIs)) in clinical medicine requires compatibility with the Digital Imaging and

Communications in Medicine (DICOM) standard. Unfortunately, WSIs usually exceed

DICOM image object size limit, making it impossible to store and exchange them in

a straightforward way. Moreover, transmitting the entire DICOM image for viewing

is ineffective for WSIs. With the JPEG2000 Interactive Protocol (JPIP), WSIs can

be linked with DICOM by transmitting image data over an auxiliary connection,

apart from patient data. In this study, we explored the feasibility of using JPIP

to link JPEG2000 WSIs with a DICOM-based Picture Archiving and Communications

System (PACS). We first modified an open-source DICOM library by adding support

for JPIP as described in the existing DICOM Supplement 106. Second, the modified

library was used as a basis for a software package (JVSdicom), which provides a

proof-of-concept for a DICOM client-server system that can transmit patient data,

conventional DICOM imagery (e.g., radiological), and JPIP-linked JPEG2000 WSIs.

The software package consists of a compression application (JVSdicom Compressor)

for producing DICOM-compatible JPEG2000 WSIs, a DICOM PACS server application

(JVSdicom Server), and a DICOM PACS client application (JVSdicom Workstation).

JVSdicom is available for free from our Web site ( http://jvsmicroscope.uta.fi/

), which also features a public JVSdicom Server, containing example X-ray images

and histopathology WSIs of breast cancer cases. The software developed indicates

that JPEG2000 and JPIP provide a well-working solution for linking WSIs with

DICOM, requiring only minor modifications to current DICOM standard

specification.

DOI: 10.1007/s10278-009-9200-1

PMCID: PMC2896636

PMID: 19415383 [Indexed for MEDLINE]

1891. BMC Bioinformatics. 2010 Jul 16;11:381. doi: 10.1186/1471-2105-11-381.

EPSVR and EPMeta: prediction of antigenic epitopes using support vector

regression and multiple server results.

Liang S(1), Zheng D, Standley DM, Yao B, Zacharias M, Zhang C.

Author information:

(1)Jacobs University Bremen, Germany.

BACKGROUND: Accurate prediction of antigenic epitopes is important for

immunologic research and medical applications, but it is still an open problem in

bioinformatics. The case for discontinuous epitopes is even worse - currently

there are only a few discontinuous epitope prediction servers available, though

discontinuous peptides constitute the majority of all B-cell antigenic epitopes.

The small number of structures for antigen-antibody complexes limits the

development of reliable discontinuous epitope prediction methods and an unbiased

benchmark to evaluate developed methods.

RESULTS: In this work, we present two novel server applications for discontinuous

epitope prediction: EPSVR and EPMeta, where EPMeta is a meta server. EPSVR,

EPMeta, and datasets are available at http://sysbio.unl.edu/services.

CONCLUSION: The server application for discontinuous epitope prediction, EPSVR,

uses a Support Vector Regression (SVR) method to integrate six scoring terms.

Furthermore, we combined EPSVR with five existing epitope prediction servers to

construct EPMeta. All methods were benchmarked by our curated independent test

set, in which all antigens had no complex structures with the antibody, and their

epitopes were identified by various biochemical experiments. The area under the

receiver operating characteristic curve (AUC) of EPSVR was 0.597, higher than

that of any other existing single server, and EPMeta had a better performance

than any single server - with an AUC of 0.638, significantly higher than PEPITO

and Disctope (p-value < 0.05).

DOI: 10.1186/1471-2105-11-381

PMCID: PMC2910724

PMID: 20637083 [Indexed for MEDLINE]

1892. BMC Pharmacol. 2010 Jul 16;10:8. doi: 10.1186/1471-2210-10-8.

Prediction of cytochrome P450 isoform responsible for metabolizing a drug

molecule.

Mishra NK(1), Agarwal S, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India.

BACKGROUND: Different isoforms of Cytochrome P450 (CYP) metabolized different

types of substrates (or drugs molecule) and make them soluble during

biotransformation. Therefore, fate of any drug molecule depends on how they are

treated or metabolized by CYP isoform. There is a need to develop models for

predicting substrate specificity of major isoforms of P450, in order to

understand whether a given drug will be metabolized or not. This paper describes

an in-silico method for predicting the metabolizing capability of major isoforms

(e.g. CYP 3A4, 2D6, 1A2, 2C9 and 2C19).

RESULTS: All models were trained and tested on 226 approved drug molecules.

Firstly, 2392 molecular descriptors for each drug molecule were calculated using

various softwares. Secondly, best 41 descriptors were selected using general and

genetic algorithm. Thirdly, Support Vector Machine (SVM) based QSAR models were

developed using 41 best descriptors and achieved an average accuracy of 86.02%,

evaluated using fivefold cross-validation. We have also evaluated the performance

of our model on an independent dataset of 146 drug molecules and achieved average

accuracy 70.55%. In addition, SVM based models were developed using 26 Chemistry

Development Kit (CDK) molecular descriptors and achieved an average accuracy of

86.60%.

CONCLUSIONS: This study demonstrates that SVM based QSAR model can predict

substrate specificity of major CYP isoforms with high accuracy. These models can

be used to predict isoform responsible for metabolizing a drug molecule. Thus

these models can used to understand whether a molecule will be metabolized or

not. This is possible to develop highly accurate models for predicting substrate

specificity of major isoforms using CDK descriptors. A web server MetaPred has

been developed for predicting metabolizing isoform of a drug molecule

http://crdd.osdd.net/raghava/metapred/.

DOI: 10.1186/1471-2210-10-8

PMCID: PMC2912882

PMID: 20637097 [Indexed for MEDLINE]

1893. Algorithms Mol Biol. 2010 Jul 15;5:29. doi: 10.1186/1748-7188-5-29.

An automated stochastic approach to the identification of the protein specificity

determinants and functional subfamilies.

Mazin PV(1), Gelfand MS, Mironov AA, Rakhmaninova AB, Rubinov AR, Russell RB,

Kalinina OV.

Author information:

(1)Cellnetworks, University of Heidelberg, Heidelberg, Germany.

olga.kalinina@bioquant.uni-heidelberg.de.

BACKGROUND: Recent progress in sequencing and 3 D structure determination

techniques stimulated development of approaches aimed at more precise annotation

of proteins, that is, prediction of exact specificity to a ligand or, more

broadly, to a binding partner of any kind.

RESULTS: We present a method, SDPclust, for identification of protein functional

subfamilies coupled with prediction of specificity-determining positions (SDPs).

SDPclust predicts specificity in a phylogeny-independent stochastic manner, which

allows for the correct identification of the specificity for proteins that are

separated on a phylogenetic tree, but still bind the same ligand. SDPclust is

implemented as a Web-server http://bioinf.fbb.msu.ru/SDPfoxWeb/ and a stand-alone

Java application available from the website.

CONCLUSIONS: SDPclust performs a simultaneous identification of specificity

determinants and specificity groups in a statistically robust and

phylogeny-independent manner.

DOI: 10.1186/1748-7188-5-29

PMCID: PMC2914642

PMID: 20633297

1894. Bioinformatics. 2010 Jul 15;26(14):1781-2. doi: 10.1093/bioinformatics/btq286.

Epub 2010 May 30.

CisGenome Browser: a flexible tool for genomic data visualization.

Jiang H(1), Wang F, Dyer NP, Wong WH.

Author information:

(1)Department of Statistics, Stanford University, Stanford, CA 94305, USA.

jiangh@stanford.edu

SUMMARY: We present an open source, platform independent tool, called CisGenome

Browser, which can work together with any other data analysis program to serve as

a flexible component for genomic data visualization. It can also work by itself

as a standalone genome browser. By working as a light-weight web server,

CisGenome Browser is a convenient tool for data sharing between labs. It has

features that are specifically designed for ultra high-throughput sequencing data

visualization.

AVAILABILITY: http://biogibbs.stanford.edu/ approximately jiangh/browser/

DOI: 10.1093/bioinformatics/btq286

PMCID: PMC2894522

PMID: 20513664 [Indexed for MEDLINE]

1895. Bioinformatics. 2010 Jul 15;26(14):1714-22. doi: 10.1093/bioinformatics/btq267.

Epub 2010 May 26.

Prediction of protease substrates using sequence and structure features.

Barkan DT(1), Hostetter DR, Mahrus S, Pieper U, Wells JA, Craik CS, Sali A.

Author information:

(1)Graduate Group in Bioinformatics, Department of Bioengineering and Therapeutic

Sciences, University of California, San Francisco, San Francisco, CA 94158, USA.

MOTIVATION: Granzyme B (GrB) and caspases cleave specific protein substrates to

induce apoptosis in virally infected and neoplastic cells. While substrates for

both types of proteases have been determined experimentally, there are many more

yet to be discovered in humans and other metazoans. Here, we present a

bioinformatics method based on support vector machine (SVM) learning that

identifies sequence and structural features important for protease recognition of

substrate peptides and then uses these features to predict novel substrates. Our

approach can act as a convenient hypothesis generator, guiding future experiments

by high-confidence identification of peptide-protein partners.

RESULTS: The method is benchmarked on the known substrates of both protease

types, including our literature-curated GrB substrate set (GrBah). On these

benchmark sets, the method outperforms a number of other methods that consider

sequence only, predicting at a 0.87 true positive rate (TPR) and a 0.13 false

positive rate (FPR) for caspase substrates, and a 0.79 TPR and a 0.21 FPR for GrB

substrates. The method is then applied to approximately 25 000 proteins in the

human proteome to generate a ranked list of predicted substrates of each protease

type. Two of these predictions, AIF-1 and SMN1, were selected for further

experimental analysis, and each was validated as a GrB substrate.

AVAILABILITY: All predictions for both protease types are publically available at

http://salilab.org/peptide. A web server is at the same site that allows a user

to train new SVM models to make predictions for any protein that recognizes

specific oligopeptide ligands.

DOI: 10.1093/bioinformatics/btq267

PMCID: PMC2894511

PMID: 20505003 [Indexed for MEDLINE]

1896. Bioinformatics. 2010 Jul 15;26(14):1708-13. doi: 10.1093/bioinformatics/btq270.

Epub 2010 May 26.

Threshold Average Precision (TAP-k): a measure of retrieval designed for

bioinformatics.

Carroll HD(1), Kann MG, Sheetlin SL, Spouge JL.

Author information:

(1)National Center for Biotechnology Information, Bethesda, MD 20894, USA.

MOTIVATION: Since database retrieval is a fundamental operation, the measurement

of retrieval efficacy is critical to progress in bioinformatics. This article

points out some issues with current methods of measuring retrieval efficacy and

suggests some improvements. In particular, many studies have used the pooled

receiver operating characteristic for n irrelevant records (ROC(n)) score, the

area under the ROC curve (AUC) of a 'pooled' ROC curve, truncated at n irrelevant

records. Unfortunately, the pooled ROC(n) score does not faithfully reflect

actual usage of retrieval algorithms. Additionally, a pooled ROC(n) score can be

very sensitive to retrieval results from as little as a single query.

METHODS: To replace the pooled ROC(n) score, we propose the Threshold Average

Precision (TAP-k), a measure closely related to the well-known average precision

in information retrieval, but reflecting the usage of E-values in bioinformatics.

Furthermore, in addition to conditions previously given in the literature, we

introduce three new criteria that an ideal measure of retrieval efficacy should

satisfy.

RESULTS: PSI-BLAST, GLOBAL, HMMER and RPS-BLAST provided examples of using the

TAP-k and pooled ROC(n) scores to evaluate sequence retrieval algorithms. In

particular, compelling examples using real data highlight the drawbacks of the

pooled ROC(n) score, showing that it can produce evaluations skewing far from

intuitive expectations. In contrast, the TAP-k satisfies most of the criteria

desired in an ideal measure of retrieval efficacy.

AVAILABILITY AND IMPLEMENTATION: The TAP-k web server and downloadable Perl

script are freely available at

http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html.ncbi/tap/

DOI: 10.1093/bioinformatics/btq270

PMCID: PMC2894514

PMID: 20505002 [Indexed for MEDLINE]

1897. BMC Bioinformatics. 2010 Jul 14;11:376. doi: 10.1186/1471-2105-11-376.

The MetabolomeExpress Project: enabling web-based processing, analysis and

transparent dissemination of GC/MS metabolomics datasets.

Carroll AJ(1), Badger MR, Harvey Millar A.

Author information:

(1)University of Western Australia, Perth, Australia. adam.carroll@anu.edu.au

BACKGROUND: Standardization of analytical approaches and reporting methods via

community-wide collaboration can work synergistically with web-tool development

to result in rapid community-driven expansion of online data repositories

suitable for data mining and meta-analysis. In metabolomics, the inter-laboratory

reproducibility of gas-chromatography/mass-spectrometry (GC/MS) makes it an

obvious target for such development. While a number of web-tools offer access to

datasets and/or tools for raw data processing and statistical analysis, none of

these systems are currently set up to act as a public repository by easily

accepting, processing and presenting publicly submitted GC/MS metabolomics

datasets for public re-analysis.

DESCRIPTION: Here, we present MetabolomeExpress, a new File Transfer Protocol

(FTP) server and web-tool for the online storage, processing, visualisation and

statistical re-analysis of publicly submitted GC/MS metabolomics datasets. Users

may search a quality-controlled database of metabolite response statistics from

publicly submitted datasets by a number of parameters (eg. metabolite, species,

organ/biofluid etc.). Users may also perform meta-analysis comparisons of

multiple independent experiments or re-analyse public primary datasets via

user-friendly tools for t-test, principal components analysis, hierarchical

cluster analysis and correlation analysis. They may interact with chromatograms,

mass spectra and peak detection results via an integrated raw data viewer.

Researchers who register for a free account may upload (via FTP) their own data

to the server for online processing via a novel raw data processing pipeline.

CONCLUSIONS: MetabolomeExpress https://www.metabolome-express.org provides a new

opportunity for the general metabolomics community to transparently present

online the raw and processed GC/MS data underlying their metabolomics

publications. Transparent sharing of these data will allow researchers to assess

data quality and draw their own insights from published metabolomics datasets.

DOI: 10.1186/1471-2105-11-376

PMCID: PMC2912306

PMID: 20626915 [Indexed for MEDLINE]

1898. J Integr Bioinform. 2010 Jul 13;7(1). doi: 10.2390/biecoll-jib-2010-143.

Kernel based machine learning algorithm for the efficient prediction of type III

polyketide synthase family of proteins.

Mallika V(1), Sivakumar KC, Jaichand S, Soniya EV.

Author information:

(1)Plant Molecular Biology Division, Rajiv Gandhi Centre for Biotechnology,

Thycaud P O, Poojappura, Thiruvananthapuram, 695 014, Kerala, India.

mallika@rgdb.res.in

Type III Polyketide synthases (PKS) are family of proteins considered to have

significant roles in the biosynthesis of various polyketides in plants, fungi and

bacteria. As these proteins shows positive effects to human health, more

researches are going on regarding this particular protein. Developing a tool to

identify the probability of sequence being a type III polyketide synthase will

minimize the time consumption and manpower efforts. In this approach, we have

designed and implemented PKSIIIpred, a high performance prediction server for

type III PKS where the classifier is Support Vector Machines (SVMs). Based on the

limited training dataset, the tool efficiently predicts the type III PKS

superfamily of proteins with high sensitivity and specificity. The PKSIIIpred is

available at http://type3pks.in/prediction/. We expect that this tool may serve

as a useful resource for type III PKS researchers. Currently work is being

progressed for further betterment of prediction accuracy by including more

sequence features in the training dataset.

DOI: 10.2390/biecoll-jib-2010-143

PMID: 20625199 [Indexed for MEDLINE]

1899. Bioinformation. 2010 Jul 6;5(2):49-51.

xFITOM: a generic GUI tool to search for transcription factor binding sites.

Bhargava N(1), Erill I.

Author information:

(1)Department of Biological Sciences, University of Maryland Baltimore

County,1000 Hilltop Circle,Baltimore,MD 21250,USA.

Locating transcription factor binding sites in genomic sequences is a key step in

deciphering transcription networks. Currently available software for site search

is mostly server-based, limiting the range and flexibility of this type of

analysis. xFITOM is a fully customizable program for locating binding sites in

genomic sequences written in C++. Through an easy-to-use interface, xFITOM that

allows users an unprecedented degree of flexibility in site search. Among other

features,it enables users to define motifs by mixing real sites and IUPAC

consensus sequences,to search the annotated sequences of unfinished genomes and

to choose among 11 different search algorithms.AVAILABILITY: XFITOM IS AVAILABLE

FOR DOWNLOAD AT: http://research.umbc.edu/˜erill.

PMCID: PMC3039987

PMID: 21346861

1900. BMC Med Educ. 2010 Jul 6;10:52. doi: 10.1186/1472-6920-10-52.

Interactive film scenes for tutor training in problem-based learning (PBL):

dealing with difficult situations.

Bosse HM(1), Huwendiek S, Skelin S, Kirschfink M, Nikendei C.

Author information:

(1)Department of General Pediatrics, Centre of Child and Adolescent Medicine, Im

Neuenheimer Feld 430, 69120 Heidelberg, Germany.

Hans.Martin.Bosse@med.uni-heidelberg.de

BACKGROUND: In problem-based learning (PBL), tutors play an essential role in

facilitating and efficiently structuring tutorials to enable students to

construct individual cognitive networks, and have a significant impact on

students' performance in subsequent assessments. The necessity of elaborate

training to fulfil this complex role is undeniable. In the plethora of data on

PBL however, little attention has been paid to tutor training which promotes

competence in the moderation of specific difficult situations commonly

encountered in PBL tutorials.

METHODS: Major interactive obstacles arising in PBL tutorials were identified

from prior publications. Potential solutions were defined by an expert group.

Video clips were produced addressing the tutor's role and providing exemplary

solutions. These clips were embedded in a PBL tutor-training course at our

medical faculty combining PBL self-experience with a non-medical case. Trainees

provided pre- and post-intervention self-efficacy ratings regarding their

PBL-related knowledge, skills, and attitudes, as well as their acceptance and the

feasibility of integrating the video clips into PBL tutor-training (all items:

100 = completely agree, 0 = don't agree at all).

RESULTS: An interactive online tool for PBL tutor training was developed

comprising 18 video clips highlighting difficult situations in PBL tutorials to

encourage trainees to develop and formulate their own intervention strategies. In

subsequent sequences, potential interventions are presented for the specific

scenario, with a concluding discussion which addresses unresolved issues. The

tool was well accepted and considered worth the time spent on it (81.62 +/-

16.91; 62.94 +/- 16.76). Tutors considered the videos to prepare them well to

respond to specific challenges in future tutorials (75.98 +/- 19.46). The entire

training, which comprised PBL self-experience and video clips as integral

elements, improved tutor's self-efficacy with respect to dealing with problematic

situations (pre: 36.47 +/- 26.25, post: 66.99 +/- 21.01; p < .0001) and

significantly increased appreciation of PBL as a method (pre: 61.33 +/- 24.84,

post: 76.20 +/- 20.12; p < .0001).

CONCLUSIONS: The interactive tool with instructional video clips is designed to

broaden the view of future PBL tutors in terms of recognizing specific obstacles

to functional group dynamics and developing individual intervention strategies.

We show that this tool is well accepted and can be successfully integrated into

PBL tutor-training. Free access is provided to the entire tool at

http://www.medizinische-fakultaet-hd.uni-heidelberg.de/fileadmin/PBLTutorTraining

/player.swf.

DOI: 10.1186/1472-6920-10-52

PMCID: PMC2909975

PMID: 20604927 [Indexed for MEDLINE]

1901. Bioinformatics. 2010 Jul 1;26(13):1673-4. doi: 10.1093/bioinformatics/btq237.

Epub 2010 May 5.

CHOYCE: a web server for constrained homology modelling with cryoEM maps.

Rawi R(1), Whitmore L, Topf M.

Author information:

(1)Institute of Structural and Molecular Biology, Crystallography, Department of

Biological Sciences, Birkbeck College, University of London, London, UK.

reda.rawi@uni-due.de

SUMMARY: CHOYCE is a web server for homology modelling of protein components and

the fitting of those components into cryo electron microscopy (cryoEM) maps of

their assemblies. It provides an interactive approach to improving the selection

of models based on the quality of their fit into the EM map.

AVAILABILITY: http://choyce.ismb.lon.ac.uk/

CONTACT: m.topf@cryst.bbk.ac.uk; reda.rawi@uni-due.de

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq237

PMCID: PMC2887048

PMID: 20444836 [Indexed for MEDLINE]

1902. Bioinformatics. 2010 Jul 1;26(13):1644-50. doi: 10.1093/bioinformatics/btq241.

Epub 2010 May 3.

Inferring the human microRNA functional similarity and functional network based

on microRNA-associated diseases.

Wang D(1), Wang J, Lu M, Song F, Cui Q.

Author information:

(1)Department of Biomedical Informatics, Peking University Health Science Center,

38 Xueyuan Road, Beijing 100191, China.

MOTIVATION: It is popular to explore meaningful molecular targets and infer new

functions of genes through gene functional similarity measuring and gene

functional network construction. However, little work is available in this field

for microRNA (miRNA) genes due to limited miRNA functional annotations. With the

rapid accumulation of miRNAs, it is increasingly needed to uncover their

functional relationships in a systems level.

RESULTS: It is known that genes with similar functions are often associated with

similar diseases, and the relationship of different diseases can be represented

by a structure of directed acyclic graph (DAG). This is also true for miRNA

genes. Therefore, it is feasible to infer miRNA functional similarity by

measuring the similarity of their associated disease DAG. Based on the above

observations and the rapidly accumulated human miRNA-disease association data, we

presented a method to infer the pairwise functional similarity and functional

network for human miRNAs based on the structures of their disease relationships.

Comparisons showed that the calculated miRNA functional similarity is well

associated with prior knowledge of miRNA functional relationship. More

importantly, this method can also be used to predict novel miRNA biomarkers and

to infer novel potential functions or associated diseases for miRNAs. In

addition, this method can be easily extended to other species when sufficient

miRNA-associated disease data are available for specific species.

AVAILABILITY: The online tool is available at http://cmbi.bjmu.edu.cn/misim

CONTACT: cuiqinghua@hsc.pku.edu.cn

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq241

PMID: 20439255 [Indexed for MEDLINE]

1903. Curr Protoc Cytom. 2010 Jul;Chapter 10:Unit10.17. doi:

10.1002/0471142956.cy1017s53.

Web-based analysis and publication of flow cytometry experiments.

Kotecha N(1), Krutzik PO, Irish JM.

Author information:

(1)Stanford University School of Medicine, Stanford, California.

Cytobank is a Web-based application for storage, analysis, and sharing of flow

cytometry experiments. Researchers use a Web browser to log in and use a wide

range of tools developed for basic and advanced flow cytometry. In addition to

providing access to standard cytometry tools from any computer, Cytobank creates

a platform and community for developing new analysis and publication tools.

Figure layouts created on Cytobank are designed to allow transparent access to

the underlying experiment annotation and data processing steps. Since all flow

cytometry files and analysis data are stored on a central server, experiments and

figures can be viewed or edited by anyone with the proper permission, from any

computer with Internet access. Once a primary researcher has performed the

initial analysis of the data, collaborators can engage in experiment analysis and

make their own figure layouts using the gated, compensated experiment files.

Cytobank is available to the scientific community at http://www.cytobank.org.

(c) 2010 by John Wiley & Sons, Inc.

DOI: 10.1002/0471142956.cy1017s53

PMCID: PMC4208272

PMID: 20578106 [Indexed for MEDLINE]

1904. Genet Epidemiol. 2010 Jul;34(5):463-8. doi: 10.1002/gepi.20504.

Evaluating the power to discriminate between highly correlated SNPs in genetic

association studies.

Udler MS(1), Tyrer J, Easton DF.

Author information:

(1)Department of Public Health and Primary Care, University of Cambridge,

Strangeways Research Laboratory, Worts Causeway, Cambridge, United Kingdom.

Neighboring common polymorphisms are often correlated (in linkage disequilibrium

(LD)) as a result of shared ancestry. An association between a polymorphism and a

disease trait may therefore be the indirect result of a correlated functional

variant, and identifying the true causal variant(s) from an initial disease

association is a major challenge in genetic association studies. Here, we present

a method to estimate the sample size needed to discriminate between a functional

variant of a given allele frequency and effect size, and other correlated

variants. The sample size required to conduct such fine-scale mapping is

typically 1-4 times larger than required to detect the initial association.

Association studies in populations with different LD patterns can substantially

improve the power to isolate the causal variant. An online tool to perform these

calculations is available at

http://moya.srl.cam.ac.uk/ocac/FineMappingPowerCalculator.html.

(c) 2010 Wiley-Liss, Inc.

DOI: 10.1002/gepi.20504

PMID: 20583289 [Indexed for MEDLINE]

1905. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W214-20. doi:

10.1093/nar/gkq537.

The GeneMANIA prediction server: biological network integration for gene

prioritization and predicting gene function.

Warde-Farley D(1), Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M,

Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright

G, Bader GD, Morris Q.

Author information:

(1)Department of Computer Science, University of Toronto, Toronto, Ontario,

Canada.

GeneMANIA (http://www.genemania.org) is a flexible, user-friendly web interface

for generating hypotheses about gene function, analyzing gene lists and

prioritizing genes for functional assays. Given a query list, GeneMANIA extends

the list with functionally similar genes that it identifies using available

genomics and proteomics data. GeneMANIA also reports weights that indicate the

predictive value of each selected data set for the query. Six organisms are

currently supported (Arabidopsis thaliana, Caenorhabditis elegans, Drosophila

melanogaster, Mus musculus, Homo sapiens and Saccharomyces cerevisiae) and

hundreds of data sets have been collected from GEO, BioGRID, Pathway Commons and

I2D, as well as organism-specific functional genomics data sets. Users can select

arbitrary subsets of the data sets associated with an organism to perform their

analyses and can upload their own data sets to analyze. The GeneMANIA algorithm

performs as well or better than other gene function prediction methods on yeast

and mouse benchmarks. The high accuracy of the GeneMANIA prediction algorithm, an

intuitive user interface and large database make GeneMANIA a useful tool for any

biologist.

DOI: 10.1093/nar/gkq537

PMCID: PMC2896186

PMID: 20576703 [Indexed for MEDLINE]

1906. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W3-6. doi: 10.1093/nar/gkq553.

Epub 2010 Jun 11.

Providing web servers and training in Bioinformatics: 2010 update on the

Bioinformatics Links Directory.

Brazas MD(1), Yamada JT, Ouellette BF.

Author information:

(1)Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto,

Ontario, Canada.

The Links Directory at Bioinformatics.ca continues its collaboration with Nucleic

Acids Research to jointly publish and compile a freely accessible, online

collection of tools, databases and resource materials for bioinformatics and

molecular biology research. The July 2010 Web Server issue of Nucleic Acids

Research adds an additional 115 web server tools and 7 updates to the directory

at http://bioinformatics.ca/links\_directory/, bringing the total number of

servers listed close to an impressive 1500 links. The Bioinformatics Links

Directory represents an excellent community resource for locating bioinformatic

tools and databases to aid one's research, and in this context bioinformatic

education needs and initiatives are discussed. A complete list of all links

featured in this Nucleic Acids Research 2010 Web Server issue can be accessed

online at http://bioinformatics.ca/links\_directory/narweb2010/. The 2010 update

of the Bioinformatics Links Directory, which includes the Web Server list and

summaries, is also available online at the Nucleic Acids Research website,

http://nar.oxfordjournals.org/.

DOI: 10.1093/nar/gkq553

PMCID: PMC2896181

PMID: 20542914 [Indexed for MEDLINE]

1907. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W523-8. doi: 10.1093/nar/gkq528.

Epub 2010 Jun 11.

MuD: an interactive web server for the prediction of non-neutral substitutions

using protein structural data.

Wainreb G(1), Ashkenazy H, Bromberg Y, Starovolsky-Shitrit A, Haliloglu T, Ruppin

E, Avraham KB, Rost B, Ben-Tal N.

Author information:

(1)Department of Biochemistry and Molecular Biology, The George S. Wise Faculty

of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel.

Erratum in

Nucleic Acids Res. 2010 Nov 1;38(21):7869.

The discrimination between functionally neutral amino acid substitutions and

non-neutral mutations, affecting protein function, is very important for our

understanding of diseases. The rapidly growing amounts of experimental data

enable the development of computational tools to facilitate the annotation of

these substitutions. Here, we describe a Random Forests-based classifier, named

Mutation Detector (MuD) that utilizes structural and sequence-derived features to

assess the impact of a given substitution on the protein function. In its

automatic mode, MuD is comparable to alternative tools in performance. However,

the uniqueness of MuD is that user-reported protein-specific structural and

functional information can be added at run-time, thereby enhancing the prediction

accuracy further. The MuD server, available at http://mud.tau.ac.il, assigns a

reliability score to every prediction, thus offering a useful tool for the

prioritization of substitutions in proteins with an available 3D structure.

DOI: 10.1093/nar/gkq528

PMCID: PMC2896130

PMID: 20542913 [Indexed for MEDLINE]

1908. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W641-6. doi: 10.1093/nar/gkq542.

Epub 2010 Jun 11.

ParticleStats: open source software for the analysis of particle motility and

cytoskeletal polarity.

Hamilton RS(1), Parton RM, Oliveira RA, Vendra G, Ball G, Nasmyth K, Davis I.

Author information:

(1)Department of Biochemistry, University of Oxford, South Parks Road, Oxford,

OX1 3QU, UK. russell.hamilton@bioch.ox.ac.uk

The study of dynamic cellular processes in living cells is central to biology and

is particularly powerful when the motility characteristics of individual objects

within cells can be determined and analysed statistically. However, commercial

programs only offer a limited range of inflexible analysis modules and there are

currently no open source programs for extensive analysis of particle motility.

Here, we describe ParticleStats (http://www.ParticleStats.com), a web server and

open source programs, which input the X,Y coordinate positions of objects in

time, and output novel analyses, graphical plots and statistics for motile

objects. ParticleStats comprises three separate analysis programs. First,

ParticleStats:Directionality for the global analysis of polarity, for example

microtubule plus end growth in Drosophila oocytes. Second, ParticleStats:Compare

for the analysis of saltatory movement in terms of runs and pauses. This can be

applied to chromosome segregation and molecular motor-based movements. Thirdly

ParticleStats:Kymographs for the analysis of kymograph images, for example as

applied to separation of chromosomes in mitosis. These analyses have provided key

insights into molecular mechanisms that are not possible from qualitative

analysis alone and are widely applicable to many other cell biology problems.

DOI: 10.1093/nar/gkq542

PMCID: PMC2896115

PMID: 20542911 [Indexed for MEDLINE]

1909. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W299-307. doi:

10.1093/nar/gkq531. Epub 2010 Jun 11.

RegPredict: an integrated system for regulon inference in prokaryotes by

comparative genomics approach.

Novichkov PS(1), Rodionov DA, Stavrovskaya ED, Novichkova ES, Kazakov AE, Gelfand

MS, Arkin AP, Mironov AA, Dubchak I.

Author information:

(1)Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

psnovichkov@lbl.org

RegPredict web server is designed to provide comparative genomics tools for

reconstruction and analysis of microbial regulons using comparative genomics

approach. The server allows the user to rapidly generate reference sets of

regulons and regulatory motif profiles in a group of prokaryotic genomes. The new

concept of a cluster of co-regulated orthologous operons allows the user to

distribute the analysis of large regulons and to perform the comparative analysis

of multiple clusters independently. Two major workflows currently implemented in

RegPredict are: (i) regulon reconstruction for a known regulatory motif and (ii)

ab initio inference of a novel regulon using several scenarios for the generation

of starting gene sets. RegPredict provides a comprehensive collection of manually

curated positional weight matrices of regulatory motifs. It is based on genomic

sequences, ortholog and operon predictions from the MicrobesOnline. An

interactive web interface of RegPredict integrates and presents diverse genomic

and functional information about the candidate regulon members from several web

resources. RegPredict is freely accessible at http://regpredict.lbl.gov.

DOI: 10.1093/nar/gkq531

PMCID: PMC2896116

PMID: 20542910 [Indexed for MEDLINE]

1910. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W576-81. doi:

10.1093/nar/gkq535. Epub 2010 Jun 11.

CPHmodels-3.0--remote homology modeling using structure-guided sequence profiles.

Nielsen M(1), Lundegaard C, Lund O, Petersen TN.

Author information:

(1)Center for Biological Sequence Analysis, Department of systems Biology, The

Technical University of Denmark, Denmark.

CPHmodels-3.0 is a web server predicting protein 3D structure by use of single

template homology modeling. The server employs a hybrid of the scoring functions

of CPHmodels-2.0 and a novel remote homology-modeling algorithm. A query sequence

is first attempted modeled using the fast CPHmodels-2.0 profile-profile scoring

function suitable for close homology modeling. The new computational costly

remote homology-modeling algorithm is only engaged provided that no suitable PDB

template is identified in the initial search. CPHmodels-3.0 was benchmarked in

the CASP8 competition and produced models for 94% of the targets (117 out of

128), 74% were predicted as high reliability models (87 out of 117). These

achieved an average RMSD of 4.6 A when superimposed to the 3D structure. The

remaining 26% low reliably models (30 out of 117) could superimpose to the true

3D structure with an average RMSD of 9.3 A. These performance values place the

CPHmodels-3.0 method in the group of high performing 3D prediction tools. Beside

its accuracy, one of the important features of the method is its speed. For most

queries, the response time of the server is <20 min. The web server is available

at http://www.cbs.dtu.dk/services/CPHmodels/.

DOI: 10.1093/nar/gkq535

PMCID: PMC2896139

PMID: 20542909 [Indexed for MEDLINE]

1911. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W53-8. doi: 10.1093/nar/gkq522.

Epub 2010 Jun 10.

ProteinDBS v2.0: a web server for global and local protein structure search.

Shyu CR(1), Pang B, Chi PH, Zhao N, Korkin D, Xu D.

Author information:

(1)Informatics Institute, University of Missouri, Columbia, MO 65211, USA.

shyuc@missouri.edu

ProteinDBS v2.0 is a web server designed for efficient and accurate comparisons

and searches of structurally similar proteins from a large-scale database. It

provides two comparison methods, global-to-global and local-to-local, to

facilitate the searches of protein structures or substructures. ProteinDBS v2.0

applies advanced feature extraction algorithms and scalable indexing techniques

to achieve a high-running speed while preserving reasonably high precision of

structural comparison. The experimental results show that our system is able to

return results of global comparisons in seconds from a complete Protein Data Bank

(PDB) database of 152,959 protein chains and that it takes much less time to

complete local comparisons from a non-redundant database of 3276 proteins than

other accurate comparison methods. ProteinDBS v2.0 supports query by PDB protein

ID and by new structures uploaded by users. To our knowledge, this is the only

search engine that can simultaneously support global and local comparisons.

ProteinDBS v2.0 is a useful tool to investigate functional or evolutional

relationships among proteins. Moreover, the common substructures identified by

local comparison can be potentially used to assist the human curation process in

discovering new domains or folds from the ever-growing protein structure

databases. The system is hosted at http://ProteinDBS.rnet.missouri.edu.

DOI: 10.1093/nar/gkq522

PMCID: PMC2896110

PMID: 20538653 [Indexed for MEDLINE]

1912. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W221-7. doi: 10.1093/nar/gkq520.

Epub 2010 Jun 10.

SoRT2: a tool for sorting genomes and reconstructing phylogenetic trees by

reversals, generalized transpositions and translocations.

Huang YL(1), Huang CC, Tang CY, Lu CL.

Author information:

(1)Department of Computer Science, National Tsing Hua University, Taiwan, R.O.C.

SoRT(2) is a web server that allows the user to perform genome rearrangement

analysis involving reversals, generalized transpositions and translocations

(including fusions and fissions), and infer phylogenetic trees of genomes being

considered based on their pairwise genome rearrangement distances. It takes as

input two or more linear/circular multi-chromosomal gene (or synteny block)

orders in FASTA-like format. When the input is two genomes, SoRT(2) will quickly

calculate their rearrangement distance, as well as a corresponding optimal

scenario by highlighting the genes involved in each rearrangement operation. In

the case of multiple genomes, SoRT(2) will also construct phylogenetic trees of

these genomes based on a matrix of their pairwise rearrangement distances using

distance-based approaches, such as neighbor-joining (NJ), unweighted pair group

method with arithmetic mean (UPGMA) and Fitch-Margoliash (FM) methods. In

addition, if the function of computing jackknife support values is selected,

SoRT(2) will further perform the jackknife analysis to evaluate statistical

reliability of the constructed NJ, UPGMA and FM trees. SoRT(2) is available

online at http://bioalgorithm.life.nctu.edu.tw/SORT2/.

DOI: 10.1093/nar/gkq520

PMCID: PMC2896082

PMID: 20538651 [Indexed for MEDLINE]

1913. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W503-7. doi: 10.1093/nar/gkq514.

Epub 2010 Jun 8.

DBCP: a web server for disulfide bonding connectivity pattern prediction without

the prior knowledge of the bonding state of cysteines.

Lin HH(1), Tseng LY.

Author information:

(1)Department of Applied Mathematics, National Chung Hsing University, Taiwan,

ROC.

The proper prediction of the location of disulfide bridges is efficient in

helping to solve the protein folding problem. Most of the previous works on the

prediction of disulfide connectivity pattern use the prior knowledge of the

bonding state of cysteines. The DBCP web server provides prediction of disulfide

bonding connectivity pattern without the prior knowledge of the bonding state of

cysteines. The method used in this server improves the accuracy of disulfide

connectivity pattern prediction (Q(p)) over the previous studies reported in the

literature. This DBCP server can be accessed at http://120.107.8.16/dbcp or

http://140.120.14.136/dbcp.

DOI: 10.1093/nar/gkq514

PMCID: PMC2896133

PMID: 20530534 [Indexed for MEDLINE]

1914. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W321-5. doi: 10.1093/nar/gkq517.

Epub 2010 Jun 6.

chipD: a web tool to design oligonucleotide probes for high-density tiling

arrays.

Dufour YS(1), Wesenberg GE, Tritt AJ, Glasner JD, Perna NT, Mitchell JC, Donohue

TJ.

Author information:

(1)Department of Bacteriology, University of Wisconsin, Madison, WI 53706, USA.

ydufour@wisc.edu

chipD is a web server that facilitates design of DNA oligonucleotide probes for

high-density tiling arrays, which can be used in a number of genomic applications

such as ChIP-chip or gene-expression profiling. The server implements a probe

selection algorithm that takes as an input, in addition to the target sequences,

a set of parameters that allow probe design to be tailored to specific

applications, protocols or the array manufacturer's requirements. The algorithm

optimizes probes to meet three objectives: (i) probes should be specific; (ii)

probes should have similar thermodynamic properties; and (iii) the target

sequence coverage should be homogeneous and avoid significant gaps. The output

provides in a text format, the list of probe sequences with their genomic

locations, targeted strands and hybridization characteristics. chipD has been

used successfully to design tiling arrays for bacteria and yeast. chipD is

available at http://chipd.uwbacter.org/.

DOI: 10.1093/nar/gkq517

PMCID: PMC2896189

PMID: 20529880 [Indexed for MEDLINE]

1915. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W293-8. doi: 10.1093/nar/gkq523.

Epub 2010 Jun 6.

SeLOX--a locus of recombination site search tool for the detection and directed

evolution of site-specific recombination systems.

Surendranath V(1), Chusainow J, Hauber J, Buchholz F, Habermann BH.

Author information:

(1)Max Planck Institute for the Molecular Cell Biology and Genetics, Dresden,

Germany.

Site-specific recombinases have become a resourceful tool for genome engineering,

allowing sophisticated in vivo DNA modifications and rearrangements, including

the precise removal of integrated retroviruses from host genomes. In a recent

study, a mutant form of Cre recombinase has been used to excise the provirus of a

specific HIV-1 strain from the human genome. To achieve provirus excision, the

Cre recombinase had to be evolved to recombine an asymmetric locus of

recombination (lox)-like sequence present in the long terminal repeat (LTR)

regions of a HIV-1 strain. One pre-requisite for this type of work is the

identification of degenerate lox-like sites in genomic sequences. Given their

nature-two inverted repeats flanking a spacer of variable length-existing search

tools like BLAST or RepeatMasker perform poorly. To address this lack of

available algorithms, we have developed the web-server SeLOX, which can identify

degenerate lox-like sites within genomic sequences. SeLOX calculates a position

weight matrix based on lox-like sequences, which is used to search genomic

sequences. For computational efficiency, we transform sequences into binary

space, which allows us to use a bit-wise AND Boolean operator for comparisons.

Next to finding lox-like sites for Cre type recombinases in HIV LTR sequences, we

have used SeLOX to identify lox-like sites in HIV LTRs for six yeast

recombinases. We finally demonstrate the general usefulness of SeLOX in

identifying lox-like sequences in large genomes by searching Cre type

recombination sites in the entire human genome. SeLOX is freely available at

http://selox.mpi-cbg.de/cgi-bin/selox/index.

DOI: 10.1093/nar/gkq523

PMCID: PMC2896191

PMID: 20529878 [Indexed for MEDLINE]

1916. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W724-31. doi:

10.1093/nar/gkq503. Epub 2010 Jun 6.

Opal web services for biomedical applications.

Ren J(1), Williams N, Clementi L, Krishnan S, Li WW.

Author information:

(1)National Biomedical Computation Resource, University of California, San Diego,

9500 Gilman Drive, La Jolla, CA 92093, USA. j2ren@ucsd.edu

Biomedical applications have become increasingly complex, and they often require

large-scale high-performance computing resources with a large number of

processors and memory. The complexity of application deployment and the advances

in cluster, grid and cloud computing require new modes of support for biomedical

research. Scientific Software as a Service (sSaaS) enables scalable and

transparent access to biomedical applications through simple standards-based Web

interfaces. Towards this end, we built a production web server

(http://ws.nbcr.net) in August 2007 to support the bioinformatics application

called MEME. The server has grown since to include docking analysis with AutoDock

and AutoDock Vina, electrostatic calculations using PDB2PQR and APBS, and

off-target analysis using SMAP. All the applications on the servers are powered

by Opal, a toolkit that allows users to wrap scientific applications easily as

web services without any modification to the scientific codes, by writing simple

XML configuration files. Opal allows both web forms-based access and programmatic

access of all our applications. The Opal toolkit currently supports SOAP-based

Web service access to a number of popular applications from the National

Biomedical Computation Resource (NBCR) and affiliated collaborative and service

projects. In addition, Opal's programmatic access capability allows our

applications to be accessed through many workflow tools, including Vision,

Kepler, Nimrod/K and VisTrails. From mid-August 2007 to the end of 2009, we have

successfully executed 239,814 jobs. The number of successfully executed jobs more

than doubled from 205 to 411 per day between 2008 and 2009. The Opal-enabled

service model is useful for a wide range of applications. It provides for

interoperation with other applications with Web Service interfaces, and allows

application developers to focus on the scientific tool and workflow development.

Web server availability: http://ws.nbcr.net.

DOI: 10.1093/nar/gkq503

PMCID: PMC2896135

PMID: 20529877 [Indexed for MEDLINE]

1917. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W201-9. doi: 10.1093/nar/gkq513.

Epub 2010 Jun 6.

SPOT: a web-based tool for using biological databases to prioritize SNPs after a

genome-wide association study.

Saccone SF(1), Bolze R, Thomas P, Quan J, Mehta G, Deelman E, Tischfield JA, Rice

JP.

Author information:

(1)Department of Psychiatry, Washington University, St Louis, MO, USA.

ssaccone@wustl.edu

SPOT (http://spot.cgsmd.isi.edu), the SNP prioritization online tool, is a web

site for integrating biological databases into the prioritization of single

nucleotide polymorphisms (SNPs) for further study after a genome-wide association

study (GWAS). Typically, the next step after a GWAS is to genotype the top

signals in an independent replication sample. Investigators will often

incorporate information from biological databases so that biologically relevant

SNPs, such as those in genes related to the phenotype or with potentially

non-neutral effects on gene expression such as a splice sites, are given higher

priority. We recently introduced the genomic information network (GIN) method for

systematically implementing this kind of strategy. The SPOT web site allows users

to upload a list of SNPs and GWAS P-values and returns a prioritized list of SNPs

using the GIN method. Users can specify candidate genes or genomic regions with

custom levels of prioritization. The results can be downloaded or viewed in the

browser where users can interactively explore the details of each SNP, including

graphical representations of the GIN method. For investigators interested in

incorporating biological databases into a post-GWAS SNP selection strategy, the

SPOT web tool is an easily implemented and flexible solution.

DOI: 10.1093/nar/gkq513

PMCID: PMC2896195

PMID: 20529875 [Indexed for MEDLINE]

1918. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W671-6. doi: 10.1093/nar/gkq497.

Epub 2010 Jun 4.

MOWServ: a web client for integration of bioinformatic resources.

Ramírez S(1), Muñoz-Mérida A, Karlsson J, García M, Pérez-Pulido AJ, Claros MG,

Trelles O.

Author information:

(1)Departamento Arquitectura de Computadores, Escuela Técnica Superior de

Ingeniería Informática, Universidad de Málaga, Málaga, Spain.

The productivity of any scientist is affected by cumbersome, tedious and

time-consuming tasks that try to make the heterogeneous web services compatible

so that they can be useful in their research. MOWServ, the bioinformatic platform

offered by the Spanish National Institute of Bioinformatics, was released to

provide integrated access to databases and analytical tools. Since its release,

the number of available services has grown dramatically, and it has become one of

the main contributors of registered services in the EMBRACE Biocatalogue. The

ontology that enables most of the web-service compatibility has been curated,

improved and extended. The service discovery has been greatly enhanced by

Magallanes software and biodataSF. User data are securely stored on the main

server by an authentication protocol that enables the monitoring of current or

already-finished user's tasks, as well as the pipelining of successive data

processing services. The BioMoby standard has been greatly extended with the new

features included in the MOWServ, such as management of additional information

(metadata such as extended descriptions, keywords and datafile examples), a

qualified registry, error handling, asynchronous services and service

replication. All of them have increased the MOWServ service quality, usability

and robustness. MOWServ is available at http://www.inab.org/MOWServ/ and has a

mirror at http://www.bitlab-es.com/MOWServ/.

DOI: 10.1093/nar/gkq497

PMCID: PMC2896175

PMID: 20525794 [Indexed for MEDLINE]

1919. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W407-11. doi:

10.1093/nar/gkq502. Epub 2010 Jun 4.

ANCHOR: a web server and database for analysis of protein-protein interaction

binding pockets for drug discovery.

Meireles LM(1), Dömling AS, Camacho CJ.

Author information:

(1)Department of Computational Biology, University of Pittsburgh, Pittsburgh, PA

15261, USA.

ANCHOR is a web-based tool whose aim is to facilitate the analysis of

protein-protein interfaces with regard to its suitability for small molecule drug

design. To this end, ANCHOR exploits the so-called anchor residues, i.e. amino

acid side-chains deeply buried at protein-protein interfaces, to indicate

possible druggable pockets to be targeted by small molecules. For a given

protein-protein complex submitted by the user, ANCHOR calculates the change in

solvent accessible surface area (DeltaSASA) upon binding for each side-chain,

along with an estimate of its contribution to the binding free energy. A

Jmol-based tool allows the user to interactively visualize selected anchor

residues in their pockets as well as the stereochemical properties of the

surrounding region such as hydrogen bonding. ANCHOR includes a Protein Data Bank

(PDB) wide database of pre-computed anchor residues from more than 30,000 PDB

entries with at least two protein chains. The user can query according to amino

acids, buried area (SASA), energy or keywords related to indication areas, e.g.

oncogene or diabetes. This database provides a resource to rapidly assess

protein-protein interactions for the suitability of small molecules or fragments

with bioisostere anchor analogues as possible compounds for pharmaceutical

intervention. ANCHOR web server and database are freely available at

http://structure.pitt.edu/anchor.

DOI: 10.1093/nar/gkq502

PMCID: PMC2896143

PMID: 20525787 [Indexed for MEDLINE]

1920. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W35-40. doi: 10.1093/nar/gkq415.

Epub 2010 Jun 4.

Multi-Harmony: detecting functional specificity from sequence alignment.

Brandt BW(1), Feenstra KA, Heringa J.

Author information:

(1)Centre for Integrative Bioinformatics, VU University Amsterdam, De Boelelaan

1081A, 1081HV Amsterdam, The Netherlands.

Many protein families contain sub-families with functional specialization, such

as binding different ligands or being involved in different protein-protein

interactions. A small number of amino acids generally determine functional

specificity. The identification of these residues can aid the understanding of

protein function and help finding targets for experimental analysis. Here, we

present multi-Harmony, an interactive web sever for detecting sub-type-specific

sites in proteins starting from a multiple sequence alignment. Combining our

Sequence Harmony (SH) and multi-Relief (mR) methods in one web server allows

simultaneous analysis and comparison of specificity residues; furthermore, both

methods have been significantly improved and extended. SH has been extended to

cope with more than two sub-groups. mR has been changed from a sampling

implementation to a deterministic one, making it more consistent and user

friendly. For both methods Z-scores are reported. The multi-Harmony web server

produces a dynamic output page, which includes interactive connections to the

Jalview and Jmol applets, thereby allowing interactive analysis of the results.

Multi-Harmony is available at http://www.ibi.vu.nl/ programs/shmrwww.

DOI: 10.1093/nar/gkq415

PMCID: PMC2896201

PMID: 20525785 [Indexed for MEDLINE]

1921. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W550-4. doi: 10.1093/nar/gkq475.

Epub 2010 Jun 4.

PUDGE: a flexible, interactive server for protein structure prediction.

Norel R(1), Petrey D, Honig B.

Author information:

(1)Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical

Institute, Center for Computational Biology and Bioinformatics, Columbia

University, 1130 St. Nicholas Avenue, New York, NY 10032, USA. rn98@columbia.edu

The construction of a homology model for a protein can involve a number of

decisions requiring the integration of different sources of information and the

application of different modeling tools depending on the particular problem.

Functional information can be especially important in guiding the modeling

process, but such information is not generally integrated into modeling

pipelines. Pudge is a flexible, interactive protein structure prediction server,

which is designed with these issues in mind. By dividing the modeling into five

stages (template selection, alignment, model building, model refinement and model

evaluation) and providing various tools to visualize, analyze and compare the

results at each stage, we enable a flexible modeling strategy that can be

tailored to the needs of a given problem. Pudge is freely available at

http://wiki.c2b2.columbia.edu/honiglab\_public/index.php/Software:PUDGE.

DOI: 10.1093/nar/gkq475

PMCID: PMC2896183

PMID: 20525783 [Indexed for MEDLINE]

1922. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W657-61. doi:

10.1093/nar/gkq498. Epub 2010 Jun 4.

SMOG@ctbp: simplified deployment of structure-based models in GROMACS.

Noel JK(1), Whitford PC, Sanbonmatsu KY, Onuchic JN.

Author information:

(1)Center for Theoretical Biological Physics and Department of Physics,

University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093,

USA.

Molecular dynamics simulations with coarse-grained and/or simplified Hamiltonians

are an effective means of capturing the functionally important long-time and

large-length scale motions of proteins and RNAs. Structure-based Hamiltonians,

simplified models developed from the energy landscape theory of protein folding,

have become a standard tool for investigating biomolecular dynamics. SMOG@ctbp is

an effort to simplify the use of structure-based models. The purpose of the web

server is two fold. First, the web tool simplifies the process of implementing a

well-characterized structure-based model on a state-of-the-art, open source,

molecular dynamics package, GROMACS. Second, the tutorial-like format helps speed

the learning curve of those unfamiliar with molecular dynamics. A web tool user

is able to upload any multi-chain biomolecular system consisting of standard RNA,

DNA and amino acids in PDB format and receive as output all files necessary to

implement the model in GROMACS. Both C(alpha) and all-atom versions of the model

are available. SMOG@ctbp resides at http://smog.ucsd.edu.

DOI: 10.1093/nar/gkq498

PMCID: PMC2896113

PMID: 20525782 [Indexed for MEDLINE]

1923. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W590-4. doi: 10.1093/nar/gkq489.

Epub 2010 Jun 3.

deconSTRUCT: general purpose protein database search on the substructure level.

Zhang ZH(1), Bharatham K, Sherman WA, Mihalek I.

Author information:

(1)Bioinformatics Institute 30 Biopolis Street, #07-01 Matrix, Singapore.

deconSTRUCT webserver offers an interface to a protein database search engine,

usable for a general purpose detection of similar protein (sub)structures.

Initially, it deconstructs the query structure into its secondary structure

elements (SSEs) and reassembles the match to the target by requiring a (tunable)

degree of similarity in the direction and sequential order of SSEs. Hierarchical

organization and judicious use of the information about protein structure enables

deconSTRUCT to achieve the sensitivity and specificity of the established search

engines at orders of magnitude increased speed, without tying up irretrievably

the substructure information in the form of a hash. In a post-processing step, a

match on the level of the backbone atoms is constructed. The results presented to

the user consist of the list of the matched SSEs, the transformation matrix for

rigid superposition of the structures and several ways of visualization, both

downloadable and implemented as a web-browser plug-in. The server is available at

http://epsf.bmad.bii.a-star.edu.sg/struct\_server.html.

DOI: 10.1093/nar/gkq489

PMCID: PMC2896154

PMID: 20522512 [Indexed for MEDLINE]

1924. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W59-63. doi: 10.1093/nar/gkq487.

Epub 2010 Jun 3.

PLAST-ncRNA: Partition function Local Alignment Search Tool for non-coding RNA

sequences.

Chikkagoudar S(1), Livesay DR, Roshan U.

Author information:

(1)Department of Biostatistics and Epidemiology, University of Pennsylvania

School of Medicine, Philadelphia, Pennsylvania, USA.

Alignment-based programs are valuable tools for finding potential homologs in

genome sequences. Previously, it has been shown that partition function posterior

probabilities attuned to local alignment achieve a high accuracy in identifying

distantly similar non-coding RNA sequences that are hidden in a large genome.

Here, we present an online implementation of that alignment algorithm based on

such probabilities. Our server takes as input a query RNA sequence and a large

genome sequence, and outputs a list of hits that are above a mean posterior

probability threshold. The output is presented in a format suited to local

alignment. It can also be viewed within the PLAST alignment viewer applet that

provides a list of all hits found and highlights regions of high posterior

probability within each local alignment. The server is freely available at

http://plastrna.njit.edu.

DOI: 10.1093/nar/gkq487

PMCID: PMC2896107

PMID: 20522510 [Indexed for MEDLINE]

1925. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W286-92. doi:

10.1093/nar/gkq473. Epub 2010 Jun 3.

TFM-Explorer: mining cis-regulatory regions in genomes.

Tonon L(1), Touzet H, Varré JS.

Author information:

(1)INRIA Lille-Nord Europe, 40 av Halley, 59650 Villeneuve d'Ascq, d'Ascq Cedex,

France.

DNA-binding transcription factors (TFs) play a central role in transcription

regulation, and computational approaches that help in elucidating complex

mechanisms governing this basic biological process are of great use. In this

perspective, we present the TFM-Explorer web server that is a toolbox to identify

putative TF binding sites within a set of upstream regulatory sequences of genes

sharing some regulatory mechanisms. TFM-Explorer finds local regions showing

overrepresentation of binding sites. Accepted organisms are human, mouse, rat,

chicken and drosophila. The server employs a number of features to help users to

analyze their data: visualization of selected binding sites on genomic sequences,

and selection of cis-regulatory modules. TFM-Explorer is available at

http://bioinfo.lifl.fr/TFM.

DOI: 10.1093/nar/gkq473

PMCID: PMC2896114

PMID: 20522509 [Indexed for MEDLINE]

1926. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W424-30. doi:

10.1093/nar/gkq480. Epub 2010 Jun 2.

SiMMap: a web server for inferring site-moiety map to recognize interaction

preferences between protein pockets and compound moieties.

Chen YF(1), Hsu KC, Lin SR, Wang WC, Huang YC, Yang JM.

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu, 30050, Taiwan.

The protein-ligand interacting mechanism is essential to biological processes and

drug discovery. The SiMMap server statistically derives site-moiety map with

several anchors, which describe the relationship between the moiety preferences

and physico-chemical properties of the binding site, from the interaction

profiles between query target protein and its docked (or co-crystallized)

compounds. Each anchor includes three basic elements: a binding pocket with

conserved interacting residues, the moiety composition of query compounds and

pocket-moiety interaction type (electrostatic, hydrogen bonding or van der

Waals). We provide initial validation of the site-moiety map on three targets,

thymidine kinase, and estrogen receptors of antagonists and agonists.

Experimental results show that an anchor is often a hot spot and the site-moiety

map can help to assemble potential leads by optimal steric, hydrogen bonding and

electronic moieties. When a compound highly agrees with anchors of site-moiety

map, this compound often activates or inhibits the target protein. We believe

that the site-moiety map is useful for drug discovery and understanding

biological mechanisms. The SiMMap web server is available at

http://simfam.life.nctu.edu.tw/.

DOI: 10.1093/nar/gkq480

PMCID: PMC2896162

PMID: 20519201 [Indexed for MEDLINE]

1927. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W508-15. doi:

10.1093/nar/gkq481. Epub 2010 May 31.

Struct2Net: a web service to predict protein-protein interactions using a

structure-based approach.

Singh R(1), Park D, Xu J, Hosur R, Berger B.

Author information:

(1)Computer Science and Artificial Intelligence Laboratory, Massachusetts

Institute of Technology, Cambridge, MA, USA.

Struct2Net is a web server for predicting interactions between arbitrary protein

pairs using a structure-based approach. Prediction of protein-protein

interactions (PPIs) is a central area of interest and successful prediction would

provide leads for experiments and drug design; however, the experimental coverage

of the PPI interactome remains inadequate. We believe that Struct2Net is the

first community-wide resource to provide structure-based PPI predictions that go

beyond homology modeling. Also, most web-resources for predicting PPIs currently

rely on functional genomic data (e.g. GO annotation, gene expression, cellular

localization, etc.). Our structure-based approach is independent of such methods

and only requires the sequence information of the proteins being queried. The web

service allows multiple querying options, aimed at maximizing flexibility. For

the most commonly studied organisms (fly, human and yeast), predictions have been

pre-computed and can be retrieved almost instantaneously. For proteins from other

species, users have the option of getting a quick-but-approximate result (using

orthology over pre-computed results) or having a full-blown computation

performed. The web service is freely available at

http://struct2net.csail.mit.edu.

DOI: 10.1093/nar/gkq481

PMCID: PMC2896152

PMID: 20513650 [Indexed for MEDLINE]

1928. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W469-73. doi:

10.1093/nar/gkq406. Epub 2010 May 31.

3DLigandSite: predicting ligand-binding sites using similar structures.

Wass MN(1), Kelley LA, Sternberg MJ.

Author information:

(1)Structural Bioinformatics Group, Centre for Bioinformatics, Imperial College

London, London, SW7 2AZ, UK.

3DLigandSite is a web server for the prediction of ligand-binding sites. It is

based upon successful manual methods used in the eighth round of the Critical

Assessment of techniques for protein Structure Prediction (CASP8). 3DLigandSite

utilizes protein-structure prediction to provide structural models for proteins

that have not been solved. Ligands bound to structures similar to the query are

superimposed onto the model and used to predict the binding site. In benchmarking

against the CASP8 targets 3DLigandSite obtains a Matthew's correlation

co-efficient (MCC) of 0.64, and coverage and accuracy of 71 and 60%,

respectively, similar results to our manual performance in CASP8. In further

benchmarking using a large set of protein structures, 3DLigandSite obtains an MCC

of 0.68. The web server enables users to submit either a query sequence or

structure. Predictions are visually displayed via an interactive Jmol applet.

3DLigandSite is available for use at http://www.sbg.bio.ic.ac.uk/3dligandsite.

DOI: 10.1093/nar/gkq406

PMCID: PMC2896164

PMID: 20513649 [Indexed for MEDLINE]

1929. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W228-32. doi:

10.1093/nar/gkq476. Epub 2010 May 31.

MARQ: an online tool to mine GEO for experiments with similar or opposite gene

expression signatures.

Vazquez M(1), Nogales-Cadenas R, Arroyo J, Botías P, García R, Carazo JM, Tirado

F, Pascual-Montano A, Carmona-Saez P.

Author information:

(1)Software Engineering Department, Facultad de Informatica, Universidad

Complutense de Madrid, Madrid, Spain.

The enormous amount of data available in public gene expression repositories such

as Gene Expression Omnibus (GEO) offers an inestimable resource to explore gene

expression programs across several organisms and conditions. This information can

be used to discover experiments that induce similar or opposite gene expression

patterns to a given query, which in turn may lead to the discovery of new

relationships among diseases, drugs or pathways, as well as the generation of new

hypotheses. In this work, we present MARQ, a web-based application that allows

researchers to compare a query set of genes, e.g. a set of over- and

under-expressed genes, against a signature database built from GEO datasets for

different organisms and platforms. MARQ offers an easy-to-use and integrated

environment to mine GEO, in order to identify conditions that induce similar or

opposite gene expression patterns to a given experimental condition. MARQ also

includes additional functionalities for the exploration of the results, including

a meta-analysis pipeline to find genes that are differentially expressed across

different experiments. The application is freely available at

http://marq.dacya.ucm.es.

DOI: 10.1093/nar/gkq476

PMCID: PMC2896165

PMID: 20513648 [Indexed for MEDLINE]

1930. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W628-32. doi:

10.1093/nar/gkq484. Epub 2010 May 31.

NMR Constraints Analyser: a web-server for the graphical analysis of NMR

experimental constraints.

Heller DM(1), Giorgetti A.

Author information:

(1)Department of Biotechnology, University of Verona, Strada Le Grazie 15, Ca'

Vignal 1, 37134 Verona, Italy.

Nuclear magnetic resonance (NMR) spectroscopy together with X-ray

crystallography, are the main techniques used for the determination of

high-resolution 3D structures of biological molecules. The output of an NMR

experiment includes a set of lower and upper limits for the distances

(constraints) between pairs of atoms. If the number of constraints is high

enough, there will be a finite number of possible conformations (models) of the

macromolecule satisfying the data. Thus, the more constraints are measured, the

better defined these structures will be. The availability of a user-friendly tool

able to help in the analysis and interpretation of the number of experimental

constraints per residue, is thus of valuable importance when assessing the levels

of structure definition of NMR solved biological macromolecules, in particular,

when high-quality structures are needed in techniques such as, computational

biology approaches, site-directed mutagenesis experiments and/or drug design.

Here, we present a free publicly available web-server, i.e. NMR Constraints

Analyser, which is aimed at providing an automatic graphical analysis of the NMR

experimental constraints atom by atom. The NMR Constraints Analyser server is

available from the web-page http://molsim.sci.univr.it/constraint.

DOI: 10.1093/nar/gkq484

PMCID: PMC2896076

PMID: 20513646 [Indexed for MEDLINE]

1931. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W516-22. doi:

10.1093/nar/gkq464. Epub 2010 May 28.

PCFamily: a web server for searching homologous protein complexes.

Lo YS(1), Lin CY, Yang JM.

Author information:

(1)Institute of Bioinformatics and Systems Biology, Department of Biological

Science and Technology and Core Facility for Structural Bioinformatics, National

Chiao Tung University, Hsinchu 30050, Taiwan.

The proteins in a cell often assemble into complexes to carry out their functions

and play an essential role of biological processes. The PCFamily server

identifies template-based homologous protein complexes [called protein complex

family (PCF)] and infers functional modules of the query proteins. This server

first finds homologous structure complexes of the query using BLASTP to search

the structural template database (11,263 complexes). PCFamily then searches the

homologous complexes of the templates (query) from a complete genomic database

(Integr8 with 6,352,363 protein sequences in 2274 species). According to these

homologous complexes across multiple species, this sever infers binding models

(e.g. hydrogen-bonds and conserved amino acids in the interfaces), functional

modules, and the conserved interacting domains and Gene Ontology annotations of

the PCF. Experimental results demonstrate that the PCFamily server can be useful

for binding model visualizations and annotating the query proteins. We believe

that the server is able to provide valuable insights for determining functional

modules of biological networks across multiple species. The PCFamily sever is

available at http://pcfamily.life.nctu.edu.tw.

DOI: 10.1093/nar/gkq464

PMCID: PMC2896147

PMID: 20511590 [Indexed for MEDLINE]

1932. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W497-502. doi:

10.1093/nar/gkq477. Epub 2010 May 27.

YLoc--an interpretable web server for predicting subcellular localization.

Briesemeister S(1), Rahnenführer J, Kohlbacher O.

Author information:

(1)Division for Simulation of Biological Systems, Universität Tübingen, Tübingen,

Germany. briese@informatik.uni-tuebingen.de

Predicting subcellular localization has become a valuable alternative to

time-consuming experimental methods. Major drawbacks of many of these predictors

is their lack of interpretability and the fact that they do not provide an

estimate of the confidence of an individual prediction. We present YLoc, an

interpretable web server for predicting subcellular localization. YLoc uses

natural language to explain why a prediction was made and which biological

property of the protein was mainly responsible for it. In addition, YLoc

estimates the reliability of its own predictions. YLoc can, thus, assist in

understanding protein localization and in location engineering of proteins. The

YLoc web server is available online at www.multiloc.org/YLoc.

DOI: 10.1093/nar/gkq477

PMCID: PMC2896088

PMID: 20507917 [Indexed for MEDLINE]

1933. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W182-7. doi: 10.1093/nar/gkq441.

Epub 2010 May 27.

waviCGH: a web application for the analysis and visualization of genomic copy

number alterations.

Carro A(1), Rico D, Rueda OM, Díaz-Uriarte R, Pisano DG.

Author information:

(1)Bioinformatics Unit, Spanish National Cancer Research Centre.

waviCGH is a versatile web server for the analysis and comparison of genomic copy

number alterations in multiple samples from any species. waviCGH processes data

generated by high density SNP-arrays, array-CGH or copy-number calls generated by

any technique. waviCGH includes methods for pre-processing of the data,

segmentation, calling of gains and losses, and minimal common regions

determination over a set of experiments. The server is a user-friendly interface

to the analytical methods, with emphasis on results visualization in a genomic

context. Analysis tools are introduced to the user as the different steps to

follow in an experimental protocol. All the analysis steps generate high quality

images and tables ready to be imported into spreadsheet programs. Additionally,

for human, mouse and rat, altered regions are represented in a biological context

by mapping them into chromosomes in an integrated cytogenetic browser. waviCGH is

available at http://wavi.bioinfo.cnio.es.

DOI: 10.1093/nar/gkq441

PMCID: PMC2896163

PMID: 20507915 [Indexed for MEDLINE]

1934. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W412-6. doi: 10.1093/nar/gkq474.

Epub 2010 May 27.

PiRaNhA: a server for the computational prediction of RNA-binding residues in

protein sequences.

Murakami Y(1), Spriggs RV, Nakamura H, Jones S.

Author information:

(1)Laboratory of Protein Informatics, Research Center for Structural and

Functional Proteomics, Institute for Protein Research, Osaka University, Osaka,

Japan.

The PiRaNhA web server is a publicly available online resource that automatically

predicts the location of RNA-binding residues (RBRs) in protein sequences. The

goal of functional annotation of sequences in the field of RNA binding is to

provide predictions of high accuracy that require only small numbers of targeted

mutations for verification. The PiRaNhA server uses a support vector machine

(SVM), with position-specific scoring matrices, residue interface propensity,

predicted residue accessibility and residue hydrophobicity as features. The

server allows the submission of up to 10 protein sequences, and the predictions

for each sequence are provided on a web page and via email. The prediction

results are provided in sequence format with predicted RBRs highlighted, in text

format with the SVM threshold score indicated and as a graph which enables users

to quickly identify those residues above any specific SVM threshold. The graph

effectively enables the increase or decrease of the false positive rate. When

tested on a non-redundant data set of 42 protein sequences not used in training,

the PiRaNhA server achieved an accuracy of 85%, specificity of 90% and a Matthews

correlation coefficient of 0.41 and outperformed other publicly available

servers. The PiRaNhA prediction server is freely available at

http://www.bioinformatics.sussex.ac.uk/PIRANHA.

DOI: 10.1093/nar/gkq474

PMCID: PMC2896099

PMID: 20507911 [Indexed for MEDLINE]

1935. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W340-7. doi: 10.1093/nar/gkq483.

Epub 2010 May 27.

iPARTS: an improved tool of pairwise alignment of RNA tertiary structures.

Wang CW(1), Chen KT, Lu CL.

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu 300, Taiwan, R.O.C.

iPARTS is an improved web server for aligning two RNA 3D structures based on a

structural alphabet (SA)-based approach. In particular, we first derive a

Ramachandran-like diagram of RNAs by plotting nucleotides on a 2D axis using

their two pseudo-torsion angles eta and . Next, we apply the affinity propagation

clustering algorithm to this eta- plot to obtain an SA of 23-nt conformations. We

finally use this SA to transform RNA 3D structures into 1D sequences of SA

letters and continue to utilize classical sequence alignment methods to compare

these 1D SA-encoded sequences and determine their structural similarities. iPARTS

takes as input two RNA 3D structures in the PDB format and outputs their global

alignment (for determining overall structural similarity), semiglobal alignments

(for detecting structural motifs or substructures), local alignments (for finding

locally similar substructures) and normalized local structural alignments (for

identifying more similar local substructures without non-similar internal

fragments), with graphical display that allows the user to visually view, rotate

and enlarge the superposition of aligned RNA 3D structures. iPARTS is now

available online at http://bioalgorithm.life.nctu.edu.tw/iPARTS/.

DOI: 10.1093/nar/gkq483

PMCID: PMC2896121

PMID: 20507908 [Indexed for MEDLINE]

1936. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W165-74. doi:

10.1093/nar/gkq472. Epub 2010 May 27.

PhenoHM: human-mouse comparative phenome-genome server.

Sardana D(1), Vasa S, Vepachedu N, Chen J, Gudivada RC, Aronow BJ, Jegga AG.

Author information:

(1)Department of Computer Science, University of Cincinnati, Cincinnati, OH, USA.

PhenoHM is a human-mouse comparative phenome-genome server that facilitates

cross-species identification of genes associated with orthologous phenotypes

(http://phenome.cchmc.org; full open access, login not required). Combining and

extrapolating the knowledge about the roles of individual gene functions in the

determination of phenotype across multiple organisms improves our understanding

of gene function in normal and perturbed states and offers the opportunity to

complement biologically the rapidly expanding strategies in comparative genomics.

The Mammalian Phenotype Ontology (MPO), a structured vocabulary of phenotype

terms that leverages observations encompassing the consequences of mouse gene

knockout studies, is a principal component of mouse phenotype knowledge source.

On the other hand, the Unified Medical Language System (UMLS) is a composite

collection of various human-centered biomedical terminologies. In the present

study, we mapped terms reciprocally from the MPO to human disease concepts such

as clinical findings from the UMLS and clinical phenotypes from the Online

Mendelian Inheritance in Man knowledgebase. By cross-mapping mouse-human

phenotype terms, extracting implicated genes and extrapolating phenotype-gene

associations between species PhenoHM provides a resource that enables rapid

identification of genes that trigger similar outcomes in human and mouse and

facilitates identification of potentially novel disease causal genes. The PhenoHM

server can be accessed freely at http://phenome.cchmc.org.

DOI: 10.1093/nar/gkq472

PMCID: PMC2896149

PMID: 20507906 [Indexed for MEDLINE]

1937. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W540-4. doi: 10.1093/nar/gkq461.

Epub 2010 May 27.

FoXS: a web server for rapid computation and fitting of SAXS profiles.

Schneidman-Duhovny D(1), Hammel M, Sali A.

Author information:

(1)Department of Bioengineering and Therapeutic Sciences, University of

California at San Francisco, CA 94158, USA.

Small angle X-ray scattering (SAXS) is an increasingly common technique for

low-resolution structural characterization of molecules in solution. SAXS

experiment determines the scattering intensity of a molecule as a function of

spatial frequency, termed SAXS profile. SAXS profiles can contribute to many

applications, such as comparing a conformation in solution with the corresponding

X-ray structure, modeling a flexible or multi-modular protein, and assembling a

macromolecular complex from its subunits. These applications require rapid

computation of a SAXS profile from a molecular structure. FoXS (Fast X-Ray

Scattering) is a rapid method for computing a SAXS profile of a given structure

and for matching of the computed and experimental profiles. Here, we describe the

interface and capabilities of the FoXS web server (http://salilab.org/foxs).

DOI: 10.1093/nar/gkq461

PMCID: PMC2896111

PMID: 20507903 [Indexed for MEDLINE]

1938. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W662-6. doi: 10.1093/nar/gkq445.

Epub 2010 May 26.

POLYVIEW-MM: web-based platform for animation and analysis of molecular

simulations.

Porollo A(1), Meller J.

Author information:

(1)Department of Environmental Health, University of Cincinnati, Cincinnati, OH,

USA. alexey.porollo@uc.edu

Molecular simulations offer important mechanistic and functional clues in studies

of proteins and other macromolecules. However, interpreting the results of such

simulations increasingly requires tools that can combine information from

multiple structural databases and other web resources, and provide highly

integrated and versatile analysis tools. Here, we present a new web server that

integrates high-quality animation of molecular motion (MM) with structural and

functional analysis of macromolecules. The new tool, dubbed POLYVIEW-MM, enables

animation of trajectories generated by molecular dynamics and related simulation

techniques, as well as visualization of alternative conformers, e.g. obtained as

a result of protein structure prediction methods or small molecule docking. To

facilitate structural analysis, POLYVIEW-MM combines interactive view and

analysis of conformational changes using Jmol and its tailored extensions,

publication quality animation using PyMol, and customizable 2D summary plots that

provide an overview of MM, e.g. in terms of changes in secondary structure states

and relative solvent accessibility of individual residues in proteins.

Furthermore, POLYVIEW-MM integrates visualization with various structural

annotations, including automated mapping of known inter-action sites from

structural homologs, mapping of cavities and ligand binding sites, transmembrane

regions and protein domains. URL: http://polyview.cchmc.org/conform.html.

DOI: 10.1093/nar/gkq445

PMCID: PMC2896192

PMID: 20504857 [Indexed for MEDLINE]

1939. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W436-40. doi:

10.1093/nar/gkq479. Epub 2010 May 26.

ProBiS: a web server for detection of structurally similar protein binding sites.

Konc J(1), Janezic D.

Author information:

(1)National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia.

A web server, ProBiS, freely available at http://probis.cmm.ki.si, is presented.

This provides access to the program ProBiS (Protein Binding Sites), which detects

protein binding sites based on local structural alignments. Detailed instructions

and user guidelines for use of ProBiS are available at the server under 'HELP'

and selected examples are provided under 'EXAMPLES'.

DOI: 10.1093/nar/gkq479

PMCID: PMC2896105

PMID: 20504855 [Indexed for MEDLINE]

1940. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W368-72. doi:

10.1093/nar/gkq432. Epub 2010 May 25.

CyloFold: secondary structure prediction including pseudoknots.

Bindewald E(1), Kluth T, Shapiro BA.

Author information:

(1)Basic Science Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD

21702, USA.

Computational RNA secondary structure prediction approaches differ by the way RNA

pseudoknot interactions are handled. For reasons of computational efficiency,

most approaches only allow a limited class of pseudoknot interactions or are not

considering them at all. Here we present a computational method for RNA secondary

structure prediction that is not restricted in terms of pseudoknot complexity.

The approach is based on simulating a folding process in a coarse-grained manner

by choosing helices based on established energy rules. The steric feasibility of

the chosen set of helices is checked during the folding process using a highly

coarse-grained 3D model of the RNA structures. Using two data sets of 26 and 241

RNA sequences we find that this approach is competitive compared to the existing

RNA secondary structure prediction programs pknotsRG, HotKnots and UnaFold. The

key advantages of the new method are that there is no algorithmic restriction in

terms of pseudoknot complexity and a test is made for steric

feasibility.AVAILABILITY: The program is available as web server at the site:

http://cylofold.abcc.ncifcrf.gov.

DOI: 10.1093/nar/gkq432

PMCID: PMC2896150

PMID: 20501603 [Indexed for MEDLINE]

1941. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W378-84. doi:

10.1093/nar/gkq431. Epub 2010 May 25.

WebPrInSeS: automated full-length clone sequence identification and verification

using high-throughput sequencing data.

Massouras A(1), Decouttere F, Hens K, Deplancke B.

Author information:

(1)Laboratory of Systems Biology and Genetics, Institute of Bioengineering,

School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015

Lausanne, Switzerland.

High-throughput sequencing (HTS) is revolutionizing our ability to obtain cheap,

fast and reliable sequence information. Many experimental approaches are expected

to benefit from the incorporation of such sequencing features in their pipeline.

Consequently, software tools that facilitate such an incorporation should be of

great interest. In this context, we developed WebPrInSeS, a web server tool

allowing automated full-length clone sequence identification and verification

using HTS data. WebPrInSeS encompasses two separate software applications. The

first is WebPrInSeS-C which performs automated sequence verification of

user-defined open-reading frame (ORF) clone libraries. The second is

WebPrInSeS-E, which identifies positive hits in cDNA or ORF-based library

screening experiments such as yeast one- or two-hybrid assays. Both tools perform

de novo assembly using HTS data from any of the three major sequencing platforms.

Thus, WebPrInSeS provides a highly integrated, cost-effective and efficient way

to sequence-verify or identify clones of interest. WebPrInSeS is available at

http://webprinses.epfl.ch/ and is open to all users.

DOI: 10.1093/nar/gkq431

PMCID: PMC2896179

PMID: 20501601 [Indexed for MEDLINE]

1942. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W281-5. doi: 10.1093/nar/gkq444.

Epub 2010 May 25.

SFmap: a web server for motif analysis and prediction of splicing factor binding

sites.

Paz I(1), Akerman M, Dror I, Kosti I, Mandel-Gutfreund Y.

Author information:

(1)Faculty of Biology, Technion-Israel Institute of Technology, Haifa 32000,

Israel.

Alternative splicing (AS) is a post-transcriptional process considered to be

responsible for the huge diversity of proteins in higher eukaryotes. AS events

are regulated by different splicing factors (SFs) that bind to sequence elements

on the RNA. SFmap is a web server for predicting putative SF binding sites in

genomic data (http://sfmap.technion.ac.il). SFmap implements the COS(WR)

algorithm, which computes similarity scores for a given regulatory motif based on

information derived from its sequence environment and its evolutionary

conservation. Input for SFmap is a human genomic sequence or a list of sequences

in FASTA format that can either be uploaded from a file or pasted into a window.

SFmap searches within a given sequence for significant hits of binding motifs

that are either stored in our database or defined by the user. SFmap results are

provided both as a text file and as a graphical web interface.

DOI: 10.1093/nar/gkq444

PMCID: PMC2896136

PMID: 20501600 [Indexed for MEDLINE]

1943. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W534-9. doi: 10.1093/nar/gkq440.

Epub 2010 May 23.

SLiMFinder: a web server to find novel, significantly over-represented, short

protein motifs.

Davey NE(1), Haslam NJ, Shields DC, Edwards RJ.

Author information:

(1)Structural and Computational Biology Unit, European Molecular Biology

Laboratory, 69117 Heidelberg, Germany.

Short, linear motifs (SLiMs) play a critical role in many biological processes,

particularly in protein-protein interactions. The Short, Linear Motif Finder

(SLiMFinder) web server is a de novo motif discovery tool that identifies

statistically over-represented motifs in a set of protein sequences, accounting

for the evolutionary relationships between them. Motifs are returned with an

intuitive P-value that greatly reduces the problem of false positives and is

accessible to biologists of all disciplines. Input can be uploaded by the user or

extracted directly from UniProt. Numerous masking options give the user great

control over the contextual information to be included in the analyses. The

SLiMFinder server combines these with user-friendly output and visualizations of

motif context to allow the user to quickly gain insight into the validity of a

putatively functional motif. These visualizations include alignments of motif

occurrences, alignments of motifs and their homologues and a visual schematic of

the top-ranked motifs. Returned motifs can also be compared with known SLiMs from

the literature using CompariMotif. All results are available for download. The

SLiMFinder server is available at: http://bioware.ucd.ie/slimfinder.html.

DOI: 10.1093/nar/gkq440

PMCID: PMC2896084

PMID: 20497999 [Indexed for MEDLINE]

1944. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W23-8. doi: 10.1093/nar/gkq443.

Epub 2010 May 23.

GUIDANCE: a web server for assessing alignment confidence scores.

Penn O(1), Privman E, Ashkenazy H, Landan G, Graur D, Pupko T.

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

Evaluating the accuracy of multiple sequence alignment (MSA) is critical for

virtually every comparative sequence analysis that uses an MSA as input. Here we

present the GUIDANCE web-server, a user-friendly, open access tool for the

identification of unreliable alignment regions. The web-server accepts as input a

set of unaligned sequences. The server aligns the sequences and provides a simple

graphic visualization of the confidence score of each column, residue and

sequence of an alignment, using a color-coding scheme. The method is generic and

the user is allowed to choose the alignment algorithm (ClustalW, MAFFT and PRANK

are supported) as well as any type of molecular sequences (nucleotide, protein or

codon sequences). The server implements two different algorithms for evaluating

confidence scores: (i) the heads-or-tails (HoT) method, which measures alignment

uncertainty due to co-optimal solutions; (ii) the GUIDANCE method, which measures

the robustness of the alignment to guide-tree uncertainty. The server projects

the confidence scores onto the MSA and points to columns and sequences that are

unreliably aligned. These can be automatically removed in preparation for

downstream analyses. GUIDANCE is freely available for use at

http://guidance.tau.ac.il.

DOI: 10.1093/nar/gkq443

PMCID: PMC2896199

PMID: 20497997 [Indexed for MEDLINE]

1945. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W313-20. doi:

10.1093/nar/gkq425. Epub 2010 May 23.

RJPrimers: unique transposable element insertion junction discovery and PCR

primer design for marker development.

You FM(1), Wanjugi H, Huo N, Lazo GR, Luo MC, Anderson OD, Dvorak J, Gu YQ.

Author information:

(1)Department of Plant Sciences, University of California, Davis, CA 95616, USA.

Transposable elements (TE) exist in the genomes of nearly all eukaryotes. TE

mobilization through 'cut-and-paste' or 'copy-and-paste' mechanisms causes their

insertions into other repetitive sequences, gene loci and other DNA. An insertion

of a TE commonly creates a unique TE junction in the genome. TE junctions are

also randomly distributed along chromosomes and therefore useful for genome-wide

marker development. Several TE-based marker systems have been developed and

applied to genetic diversity assays, and to genetic and physical mapping. A

software tool 'RJPrimers' reported here allows for accurate identification of

unique repeat junctions using BLASTN against annotated repeat databases and a

repeat junction finding algorithm, and then for fully automated high-throughput

repeat junction-based primer design using Primer3 and BatchPrimer3. The software

was tested using the rice genome and genomic sequences of Aegilops tauschii. Over

90% of repeat junction primers designed by RJPrimers were unique. At least one

RJM marker per 10 Kb sequence of A. tauschii was expected with an estimate of

over 0.45 million such markers in a genome of 4.02 Gb, providing an almost

unlimited source of molecular markers for mapping large and complex genomes. A

web-based server and a command line-based pipeline for RJPrimers are both

available at http://wheat.pw.usda.gov/demos/RJPrimers/.

DOI: 10.1093/nar/gkq425

PMCID: PMC2896120

PMID: 20497996 [Indexed for MEDLINE]

1946. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W188-93. doi:

10.1093/nar/gkq430. Epub 2010 May 21.

THREaD Mapper Studio: a novel, visual web server for the estimation of genetic

linkage maps.

Cheema J(1), Ellis TH, Dicks J.

Author information:

(1)Department of Computational and Systems Biology, John Innes Centre, Norwich

Research Park, Colney, Norwich, NR4 7UH, UK.

The estimation of genetic linkage maps is a key component in plant and animal

research, providing both an indication of the genetic structure of an organism

and a mechanism for identifying candidate genes associated with traits of

interest. Because of this importance, several computational solutions to genetic

map estimation exist, mostly implemented as stand-alone software packages.

However, the estimation process is often largely hidden from the user.

Consequently, problems such as a program crashing may occur that leave a user

baffled. THREaD Mapper Studio (http://cbr.jic.ac.uk/threadmapper) is a new web

site that implements a novel, visual and interactive method for the estimation of

genetic linkage maps from DNA markers. The rationale behind the web site is to

make the estimation process as transparent and robust as possible, while also

allowing users to use their expert knowledge during analysis. Indeed, the 3D

visual nature of the tool allows users to spot features in a data set, such as

outlying markers and potential structural rearrangements that could cause

problems with the estimation procedure and to account for them in their analysis.

Furthermore, THREaD Mapper Studio facilitates the visual comparison of genetic

map solutions from third party software, aiding users in developing robust

solutions for their data sets.

DOI: 10.1093/nar/gkq430

PMCID: PMC2896177

PMID: 20494977 [Indexed for MEDLINE]

1947. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W109-17. doi:

10.1093/nar/gkq424. Epub 2010 May 21.

Discovering causal signaling pathways through gene-expression patterns.

Parikh JR(1), Klinger B, Xia Y, Marto JA, Blüthgen N.

Author information:

(1)Bioinformatics Program, Boston University, Boston, MA 02115, USA.

High-throughput gene-expression studies result in lists of differentially

expressed genes. Most current meta-analyses of these gene lists include searching

for significant membership of the translated proteins in various signaling

pathways. However, such membership enrichment algorithms do not provide insight

into which pathways caused the genes to be differentially expressed in the first

place. Here, we present an intuitive approach for discovering upstream signaling

pathways responsible for regulating these differentially expressed genes. We

identify consistently regulated signature genes specific for signal transduction

pathways from a panel of single-pathway perturbation experiments. An algorithm

that detects overrepresentation of these signature genes in a gene group of

interest is used to infer the signaling pathway responsible for regulation. We

expose our novel resource and algorithm through a web server called SPEED:

Signaling Pathway Enrichment using Experimental Data sets. SPEED can be freely

accessed at http://speed.sys-bio.net/.

DOI: 10.1093/nar/gkq424

PMCID: PMC2896193

PMID: 20494976 [Indexed for MEDLINE]

1948. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W602-8. doi: 10.1093/nar/gkq401.

Epub 2010 May 19.

MPlot--a server to analyze and visualize tertiary structure contacts and

geometrical features of helical membrane proteins.

Rose A(1), Goede A, Hildebrand PW.

Author information:

(1)Charité, Institute of Medical Physics and Biophysics, ProteInformatics Group,

Ziegelstr. 7/9, Berlin, Germany.

MPlot is a webserver that provides a quick and easy way for structural biologists

to analyze, visualize and plot tertiary structure contacts of helical membrane

proteins. As input, experimentally determined or computationally modeled protein

structures in PDB format are required. The automatic analysis concatenates in

house tools to calculate cut-off dependent van der Waals contacts or crossing

angles of transmembrane helices with third party tools to compute main chain or

side chain hydrogen bonds or membrane planes. Moreover, MPlot allows new features

and tools to be added on a regular basis. For that purpose, MPlot was embedded in

a framework that facilitates advanced users to compose new workflows from

existing tools, or to substitute intermediate results with results from their

(own) tools. The outputs can be viewed online in a Jmol based protein viewer, or

via automatically generated scripts in PyMOL. For further illustration, the

results can be downloaded as a 2D graph, representing the spatial arrangement of

transmembrane helices true to scale. For analysis and statistics, all results can

be downloaded as text files that may serve as inputs for or as standard data to

validate the output of knowledge based tertiary structure prediction tools. URL:

http://proteinformatics.charite.de/mplot/.

DOI: 10.1093/nar/gkq401

PMCID: PMC2896131

PMID: 20484376 [Indexed for MEDLINE]

1949. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W96-102. doi:

10.1093/nar/gkq418. Epub 2010 May 19.

ToppCluster: a multiple gene list feature analyzer for comparative enrichment

clustering and network-based dissection of biological systems.

Kaimal V(1), Bardes EE, Tabar SC, Jegga AG, Aronow BJ.

Author information:

(1)Division of Biomedical Informatics, Cincinnati Children's Hospital Medical

Center, Cincinnati, OH, USA.

ToppCluster is a web server application that leverages a powerful enrichment

analysis and underlying data environment for comparative analyses of multiple

gene lists. It generates heatmaps or connectivity networks that reveal functional

features shared or specific to multiple gene lists. ToppCluster uses

hypergeometric tests to obtain list-specific feature enrichment P-values for

currently 17 categories of annotations of human-ortholog genes, and provides

user-selectable cutoffs and multiple testing correction methods to control false

discovery. Each nameable gene list represents a column input to a resulting

matrix whose rows are overrepresented features, and individual cells per-list

P-values and corresponding genes per feature. ToppCluster provides users with

choices of tabular outputs, hierarchical clustering and heatmap generation, or

the ability to interactively select features from the functional enrichment

matrix to be transformed into XGMML or GEXF network format documents for use in

Cytoscape or Gephi applications, respectively. Here, as example, we demonstrate

the ability of ToppCluster to enable identification of list-specific phenotypic

and regulatory element features (both cis-elements and 3'UTR microRNA binding

sites) among tissue-specific gene lists. ToppCluster's functionalities enable the

identification of specialized biological functions and regulatory networks and

systems biology-based dissection of biological states. ToppCluster can be

accessed freely at http://toppcluster.cchmc.org.

DOI: 10.1093/nar/gkq418

PMCID: PMC2896202

PMID: 20484371 [Indexed for MEDLINE]

1950. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W474-9. doi: 10.1093/nar/gkq407.

Epub 2010 May 18.

MetaMHC: a meta approach to predict peptides binding to MHC molecules.

Hu X(1), Zhou W, Udaka K, Mamitsuka H, Zhu S.

Author information:

(1)School of Computer Science and Shanghai Key Lab of Intelligent Information

Processing, Fudan University, Shanghai 200433, China.

As antigenic peptides binding to major histocompatibility complex (MHC) molecules

is the prerequisite of cellular immune responses, an accurate computational

predictor will be of great benefit to biologists and immunologists for

understanding the underlying mechanism of immune recognition as well as

facilitating the process of epitope mapping and vaccine design. Although various

computational approaches have been developed, recent experimental results on

benchmark data sets show that the development of improved predictors is needed,

especially for MHC Class II peptide binding. To make the most of current methods

and achieve a higher predictive performance, we developed a new web server,

MetaMHC, to integrate the outputs of leading predictors by several popular

ensemble strategies. MetaMHC consists of two components: MetaMHCI and MetaMHCII

for MHC Class I peptide and MHC Class II peptide binding predictions,

respectively. Experimental results by both cross-validation and using an

independent data set show that the ensemble approaches outperform individual

predictors, being statistically significant. MetaMHC is freely available at

http://www.biokdd.fudan.edu.cn/Service/MetaMHC.html.

DOI: 10.1093/nar/gkq407

PMCID: PMC2896142

PMID: 20483919 [Indexed for MEDLINE]

1951. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W431-5. doi: 10.1093/nar/gkq361.

Epub 2010 May 16.

NAPS: a residue-level nucleic acid-binding prediction server.

Carson MB(1), Langlois R, Lu H.

Author information:

(1)Department of Bioengineering/Bioinformatics, University of Illinois at

Chicago, Chicago, IL, USA.

Nucleic acid-binding proteins are involved in a great number of cellular

processes. Understanding the mechanisms underlying these proteins first requires

the identification of specific residues involved in nucleic acid binding.

Prediction of NA-binding residues can provide practical assistance in the

functional annotation of NA-binding proteins. Predictions can also be used to

expedite mutagenesis experiments, guiding researchers to the correct binding

residues in these proteins. Here, we present a method for the identification of

amino acid residues involved in DNA- and RNA-binding using sequence-based

attributes. The method used in this work combines the C4.5 algorithm with

bootstrap aggregation and cost-sensitive learning. Our DNA-binding model achieved

79.1% accuracy, while the RNA-binding model reached an accuracy of 73.2%. The

NAPS web server is freely available at http://proteomics.bioengr.uic.edu/NAPS.

DOI: 10.1093/nar/gkq361

PMCID: PMC2896077

PMID: 20478832 [Indexed for MEDLINE]

1952. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W262-7. doi: 10.1093/nar/gkq391.

Epub 2010 May 16.

RSSsite: a reference database and prediction tool for the identification of

cryptic Recombination Signal Sequences in human and murine genomes.

Merelli I(1), Guffanti A, Fabbri M, Cocito A, Furia L, Grazini U, Bonnal RJ,

Milanesi L, McBlane F.

Author information:

(1)Institute for Biomedical Technologies, National Research Council, via Fratelli

Cervi 93, 20090 Segrate, Milano, Italy. ivan.merelli@itb.cnr.it

Recombination signal sequences (RSSs) flanking V, D and J gene segments are

recognized and cut by the VDJ recombinase during development of B and T

lymphocytes. All RSSs are composed of seven conserved nucleotides, followed by a

spacer (containing either 12 +/- 1 or 23 +/- 1 poorly conserved nucleotides) and

a conserved nonamer. Errors in V(D)J recombination, including cleavage of cryptic

RSS outside the immunoglobulin and T cell receptor loci, are associated with

oncogenic translocations observed in some lymphoid malignancies. We present in

this paper the RSSsite web server, which is available from the address

http://www.itb.cnr.it/rss. RSSsite consists of a web-accessible database, RSSdb,

for the identification of pre-computed potential RSSs, and of the related search

tool, DnaGrab, which allows the scoring of potential RSSs in user-supplied

sequences. This latter algorithm makes use of probability models, which can be

recasted to Bayesian network, taking into account correlations between groups of

positions of a sequence, developed starting from specific reference sets of RSSs.

In validation laboratory experiments, we selected 33 predicted cryptic RSSs

(cRSSs) from 11 chromosomal regions outside the immunoglobulin and TCR loci for

functional testing.

DOI: 10.1093/nar/gkq391

PMCID: PMC2896083

PMID: 20478831 [Indexed for MEDLINE]

1953. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W529-33. doi:

10.1093/nar/gkq399. Epub 2010 May 16.

ConSurf 2010: calculating evolutionary conservation in sequence and structure of

proteins and nucleic acids.

Ashkenazy H(1), Erez E, Martz E, Pupko T, Ben-Tal N.

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

It is informative to detect highly conserved positions in proteins and nucleic

acid sequence/structure since they are often indicative of structural and/or

functional importance. ConSurf (http://consurf.tau.ac.il) and ConSeq

(http://conseq.tau.ac.il) are two well-established web servers for calculating

the evolutionary conservation of amino acid positions in proteins using an

empirical Bayesian inference, starting from protein structure and sequence,

respectively. Here, we present the new version of the ConSurf web server that

combines the two independent servers, providing an easier and more intuitive

step-by-step interface, while offering the user more flexibility during the

process. In addition, the new version of ConSurf calculates the evolutionary

rates for nucleic acid sequences. The new version is freely available at:

http://consurf.tau.ac.il/.

DOI: 10.1093/nar/gkq399

PMCID: PMC2896094

PMID: 20478830 [Indexed for MEDLINE]

1954. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W582-9. doi: 10.1093/nar/gkq383.

Epub 2010 May 16.

fpocket: online tools for protein ensemble pocket detection and tracking.

Schmidtke P(1), Le Guilloux V, Maupetit J, Tufféry P.

Author information:

(1)Departament de Fisicoquimica and Institut de Biomedicina, Facultat de

Farmacia, Universitat de Barcelona, 08028, Barcelona, Spain.

Computational small-molecule binding site detection has several important

applications in the biomedical field. Notable interests are the identification of

cavities for structure-based drug discovery or functional annotation of

structures. fpocket is a small-molecule pocket detection program, relying on the

geometric alpha-sphere theory. The fpocket web server allows: (i) candidate

pocket detection--fpocket; (ii) pocket tracking during molecular dynamics, in

order to provide insights into pocket dynamics--mdpocket; and (iii) a

transposition of mdpocket to the combined analysis of homologous

structures--hpocket. These complementary online tools allow to tackle various

questions related to the identification and annotation of functional and

allosteric sites, transient pockets and pocket preservation within evolution of

structural families. The server and documentation are freely available at

http://bioserv.rpbs.univ-paris-diderot.fr/fpocket.

DOI: 10.1093/nar/gkq383

PMCID: PMC2896101

PMID: 20478829 [Indexed for MEDLINE]

1955. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W392-7. doi: 10.1093/nar/gkq393.

Epub 2010 May 16.

mirTools: microRNA profiling and discovery based on high-throughput sequencing.

Zhu E(1), Zhao F, Xu G, Hou H, Zhou L, Li X, Sun Z, Wu J.

Author information:

(1)Institute of Genomic Medicine/Zhejiang Provincial Key Laboratory of Medical

Genetics, Wenzhou, China.

miRNAs are small, non-coding RNA that negatively regulate gene expression at

post-transcriptional level, which play crucial roles in various physiological and

pathological processes, such as development and tumorigenesis. Although deep

sequencing technologies have been applied to investigate various small RNA

transcriptomes, their computational methods are far away from maturation as

compared to microarray-based approaches. In this study, a comprehensive web

server mirTools was developed to allow researchers to comprehensively

characterize small RNA transcriptome. With the aid of mirTools, users can: (i)

filter low-quality reads and 3/5' adapters from raw sequenced data; (ii) align

large-scale short reads to the reference genome and explore their length

distribution; (iii) classify small RNA candidates into known categories, such as

known miRNAs, non-coding RNA, genomic repeats and coding sequences; (iv) provide

detailed annotation information for known miRNAs, such as miRNA/miRNA\*,

absolute/relative reads count and the most abundant tag; (v) predict novel miRNAs

that have not been characterized before; and (vi) identify differentially

expressed miRNAs between samples based on two different counting strategies:

total read tag counts and the most abundant tag counts. We believe that the

integration of multiple computational approaches in mirTools will greatly

facilitate current microRNA researches in multiple ways. mirTools can be accessed

at http://centre.bioinformatics.zj.cn/mirtools/ and http://59.79.168.90/mirtools.

DOI: 10.1093/nar/gkq393

PMCID: PMC2896132

PMID: 20478827 [Indexed for MEDLINE]

1956. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W555-62. doi:

10.1093/nar/gkq395. Epub 2010 May 16.

3V: cavity, channel and cleft volume calculator and extractor.

Voss NR(1), Gerstein M.

Author information:

(1)Department of Cell Biology, The Scripps Research Institute, CB 129, La Jolla,

CA 92037, USA.

As larger macromolecular structures become available, there is a growing need to

understand their 'internal' volumes--such as deep clefts, channels and

cavities--as these often play critical roles in their function. The 3V web server

can automatically extract and comprehensively analyze all the internal volumes

from input RNA and protein structures. It rapidly finds internal volumes by

taking the difference between two rolling-probe solvent-excluded surfaces, one

with as large as possible a probe radius and the other with a solvent radius

(typically 1.5 A for water). The outputs are volumetric representations, both as

images and downloadable files, which can be used for further analysis. The 3V

server and source code are available from http://3vee.molmovdb.org.

DOI: 10.1093/nar/gkq395

PMCID: PMC2896178

PMID: 20478824 [Indexed for MEDLINE]

1957. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W706-11. doi:

10.1093/nar/gkq386. Epub 2010 May 14.

TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web

services.

Katayama T(1), Nakao M, Takagi T.

Author information:

(1)Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1

Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan. ktym@hgc.jp

Web services have become widely used in bioinformatics analysis, but there exist

incompatibilities in interfaces and data types, which prevent users from making

full use of a combination of these services. Therefore, we have developed the

TogoWS service to provide an integrated interface with advanced features. In the

TogoWS REST (REpresentative State Transfer) API (application programming

interface), we introduce a unified access method for major database resources

through intuitive URIs that can be used to search, retrieve, parse and convert

the database entries. The TogoWS SOAP API resolves compatibility issues found on

the server and client-side SOAP implementations. The TogoWS service is freely

available at: http://togows.dbcls.jp/.

DOI: 10.1093/nar/gkq386

PMCID: PMC2896079

PMID: 20472643 [Indexed for MEDLINE]

1958. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W595-601. doi:

10.1093/nar/gkq398. Epub 2010 May 12.

GRAPE: GRaphical Abstracted Protein Explorer.

Cipriano G(1), Wesenberg G, Grim T, Phillips GN Jr, Gleicher M.

Author information:

(1)Department of Computer Sciences, University of Wisconsin-Madison Madison, WI

53715, USA. gregc@cs.wisc.edu

The region surrounding a protein, known as the surface of interaction or

molecular surface, can provide valuable insight into its function. Unfortunately,

due to the complexity of both their geometry and their surface fields, study of

these surfaces can be slow and difficult and important features may be hard to

identify. Here, we describe our GRaphical Abstracted Protein Explorer, or GRAPE,

a web server that allows users to explore abstracted representations of proteins.

These abstracted surfaces effectively reduce the level of detail of the surface

of a macromolecule, using a specialized algorithm that removes small bumps and

pockets, while preserving large-scale structural features. Scalar fields, such as

electrostatic potential and hydropathy, are smoothed to further reduce visual

complexity. This entirely new way of looking at proteins complements more

traditional views of the molecular surface. GRAPE includes a thin 3D viewer that

allows users to quickly flip back and forth between both views. Abstracted views

provide a fast way to assess both a molecule's shape and its different surface

field distributions. GRAPE is freely available at http://grape.uwbacter.org.

DOI: 10.1093/nar/gkq398

PMCID: PMC2896102

PMID: 20462864 [Indexed for MEDLINE]

1959. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W569-75. doi:

10.1093/nar/gkq369. Epub 2010 May 12.

RosettaBackrub--a web server for flexible backbone protein structure modeling and

design.

Lauck F(1), Smith CA, Friedland GF, Humphris EL, Kortemme T.

Author information:

(1)Department of Bioengineering and Therapeutic Sciences, University of

California at San Francisco, 1700 4th Street, San Francisco, CA 94158, USA.

The RosettaBackrub server (http://kortemmelab.ucsf.edu/backrub) implements the

Backrub method, derived from observations of alternative conformations in

high-resolution protein crystal structures, for flexible backbone protein

modeling. Backrub modeling is applied to three related applications using the

Rosetta program for structure prediction and design: (I) modeling of structures

of point mutations, (II) generating protein conformational ensembles and

designing sequences consistent with these conformations and (III) predicting

tolerated sequences at protein-protein interfaces. The three protocols have been

validated on experimental data. Starting from a user-provided single input

protein structure in PDB format, the server generates near-native conformational

ensembles. The predicted conformations and sequences can be used for different

applications, such as to guide mutagenesis experiments, for ensemble-docking

approaches or to generate sequence libraries for protein design.

DOI: 10.1093/nar/gkq369

PMCID: PMC2896185

PMID: 20462859 [Indexed for MEDLINE]

1960. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W633-40. doi:

10.1093/nar/gkq375. Epub 2010 May 11.

PROSESS: a protein structure evaluation suite and server.

Berjanskii M(1), Liang Y, Zhou J, Tang P, Stothard P, Zhou Y, Cruz J, MacDonell

C, Lin G, Lu P, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Canada.

PROSESS (PROtein Structure Evaluation Suite and Server) is a web server designed

to evaluate and validate protein structures generated by X-ray crystallography,

NMR spectroscopy or computational modeling. While many structure evaluation

packages have been developed over the past 20 years, PROSESS is unique in its

comprehensiveness, its capacity to evaluate X-ray, NMR and predicted structures

as well as its ability to evaluate a variety of experimental NMR data. PROSESS

integrates a variety of previously developed, well-known and thoroughly tested

methods to evaluate both global and residue specific: (i) covalent and geometric

quality; (ii) non-bonded/packing quality; (iii) torsion angle quality; (iv)

chemical shift quality and (v) NOE quality. In particular, PROSESS uses VADAR for

coordinate, packing, H-bond, secondary structure and geometric analysis, GeNMR

for calculating folding, threading and solvent energetics, ShiftX for calculating

chemical shift correlations, RCI for correlating structure mobility to chemical

shift and PREDITOR for calculating torsion angle-chemical shifts agreement.

PROSESS also incorporates several other programs including MolProbity to assess

atomic clashes, Xplor-NIH to identify and quantify NOE restraint violations and

NAMD to assess structure energetics. PROSESS produces detailed tables,

explanations, structural images and graphs that summarize the results and compare

them to values observed in high-quality or high-resolution protein structures.

Using a simplified red-amber-green coloring scheme PROSESS also alerts users

about both general and residue-specific structural problems. PROSESS is intended

to serve as a tool that can be used by structure biologists as well as database

curators to assess and validate newly determined protein structures. PROSESS is

freely available at http://www.prosess.ca.

DOI: 10.1093/nar/gkq375

PMCID: PMC2896095

PMID: 20460469 [Indexed for MEDLINE]

1961. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W360-7. doi: 10.1093/nar/gkq371.

Epub 2010 May 11.

SIREs: searching for iron-responsive elements.

Campillos M(1), Cases I, Hentze MW, Sanchez M.

Author information:

(1)European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg,

Germany.

The iron regulatory protein/iron-responsive element regulatory system plays a

crucial role in the post-transcriptional regulation of gene expression and its

disruption results in human disease. IREs are cis-acting regulatory motifs

present in mRNAs that encode proteins involved in iron metabolism. They function

as binding sites for two related trans-acting factors, namely the IRP-1 and -2.

Among cis-acting RNA regulatory elements, the IRE is one of the best

characterized. It is defined by a combination of RNA sequence and structure.

However, currently available programs to predict IREs do not show a satisfactory

level of sensitivity and fail to detect some of the functional IREs. Here, we

report an improved software for the prediction of IREs implemented as a

user-friendly web server tool. The SIREs web server uses a simple data input

interface and provides structure analysis, predicted RNA folds, folding energy

data and an overall quality flag based on properties of well characterized IREs.

Results are reported in a tabular format and as a schematic visual representation

that highlights important features of the IRE. The SIREs (Search for

iron-responsive elements) web server is freely available on the web at

http://ccbg.imppc.org/sires/index.html.

DOI: 10.1093/nar/gkq371

PMCID: PMC2896125

PMID: 20460462 [Indexed for MEDLINE]

1962. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W457-61. doi:

10.1093/nar/gkq373. Epub 2010 May 11.

FiberDock: a web server for flexible induced-fit backbone refinement in molecular

docking.

Mashiach E(1), Nussinov R, Wolfson HJ.

Author information:

(1)Blavatnik School of Computer Science, Raymond and Beverly Sackler Faculty of

Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

Protein-protein docking algorithms aim to predict the structure of a complex

given the atomic structures of the proteins that assemble it. The docking

procedure usually consists of two main steps: docking candidate generation and

their refinement. The refinement stage aims to improve the accuracy of the

candidate solutions and to identify near-native solutions among them. During

protein-protein interaction, both side chains and backbone change their

conformation. Refinement methods should model these conformational changes in

order to obtain a more accurate model of the complex. Handling protein backbone

flexibility is a major challenge for docking methodologies, since backbone

flexibility adds a huge number of degrees of freedom to the search space.

FiberDock is the first docking refinement web server, which accounts for both

backbone and side-chain flexibility. Given a set of up to 100 potential docking

candidates, FiberDock models the backbone and side-chain movements that occur

during the interaction, refines the structures and scores them according to an

energy function. The FiberDock web server is free and available with no login

requirement at http://bioinfo3d.cs.tau.ac.il/FiberDock/.

DOI: 10.1093/nar/gkq373

PMCID: PMC2896170

PMID: 20460459 [Indexed for MEDLINE]

1963. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W398-401. doi:

10.1093/nar/gkq360. Epub 2010 May 10.

CCRXP: exploring clusters of conserved residues in protein structures.

Ahmad S(1), Keskin O, Mizuguchi K, Sarai A, Nussinov R.

Author information:

(1)National Institute of Biomedical Innovation, 7-6-8, Saito-asagi, Ibaraki,

Osaka 5670085, Japan. shandar@nibio.go.jp

Conserved residues forming tightly packed clusters have been shown to be energy

hot spots in both protein-protein and protein-DNA complexes. A number of analyses

on these clusters of conserved residues (CCRs) have been reported, all pointing

to a crucial role that these clusters play in protein function, especially

protein-protein and protein-DNA interactions. However, currently there is no

publicly available tool to automatically detect such clusters. Here, we present a

web server that takes a coordinate file in PDB format as input and automatically

executes all the steps to identify CCRs in protein structures. In addition, it

calculates the structural properties of each residue and of the CCRs. We also

present statistics to show that CCRs, determined by these procedures, are

significantly enriched in 'hot spots' in protein-protein and protein-RNA

complexes, which supplements our more detailed similar results on protein-DNA

complexes. We expect that CCRXP web server will be useful in studies of protein

structures and their interactions and selecting mutagenesis targets. The web

server can be accessed at http://ccrxp.netasa.org.

DOI: 10.1093/nar/gkq360

PMCID: PMC2896124

PMID: 20457748 [Indexed for MEDLINE]

1964. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W71-7. doi: 10.1093/nar/gkq329.

Epub 2010 May 10.

MSEA: a web-based tool to identify biologically meaningful patterns in

quantitative metabolomic data.

Xia J(1), Wishart DS.

Author information:

(1)Department of Biological Sciences, University of Alberta, Edmonton, AB,

Canada.

Gene set enrichment analysis (GSEA) is a widely used technique in transcriptomic

data analysis that uses a database of predefined gene sets to rank lists of genes

from microarray studies to identify significant and coordinated changes in gene

expression data. While GSEA has been playing a significant role in understanding

transcriptomic data, no similar tools are currently available for understanding

metabolomic data. Here, we introduce a web-based server, called Metabolite Set

Enrichment Analysis (MSEA), to help researchers identify and interpret patterns

of human or mammalian metabolite concentration changes in a biologically

meaningful context. Key to the development of MSEA has been the creation of a

library of approximately 1000 predefined metabolite sets covering various

metabolic pathways, disease states, biofluids, and tissue locations. MSEA also

supports user-defined or custom metabolite sets for more specialized analysis.

MSEA offers three different enrichment analyses for metabolomic studies including

overrepresentation analysis (ORA), single sample profiling (SSP) and quantitative

enrichment analysis (QEA). ORA requires only a list of compound names, while SSP

and QEA require both compound names and compound concentrations. MSEA generates

easily understood graphs or tables embedded with hyperlinks to relevant pathway

images and disease descriptors. For non-mammalian or more specialized metabolomic

studies, MSEA allows users to provide their own metabolite sets for enrichment

analysis. The MSEA server also supports conversion between metabolite common

names, synonyms, and major database identifiers. MSEA has the potential to help

users identify obvious as well as 'subtle but coordinated' changes among a group

of related metabolites that may go undetected with conventional approaches. MSEA

is freely available at http://www.msea.ca.

DOI: 10.1093/nar/gkq329

PMCID: PMC2896187

PMID: 20457745 [Indexed for MEDLINE]

1965. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W545-9. doi: 10.1093/nar/gkq366.

Epub 2010 May 10.

Dali server: conservation mapping in 3D.

Holm L(1), Rosenström P.

Author information:

(1)Institute of Biotechnology and Department of Biosciences, University of

Helsinki, Helsinki, Finland. liisa.holm@helsinki.fi

Our web site (http://ekhidna.biocenter.helsinki.fi/dali\_server) runs the Dali

program for protein structure comparison. The web site consists of three parts:

(i) the Dali server compares newly solved structures against structures in the

Protein Data Bank (PDB), (ii) the Dali database allows browsing precomputed

structural neighbourhoods and (iii) the pairwise comparison generates suboptimal

alignments for a pair of structures. Each part has its own query form and a

common format for the results page. The inputs are either PDB identifiers or

novel structures uploaded by the user. The results pages are hyperlinked to aid

interactive analysis. The web interface is simple and easy to use. The key

purpose of interactive analysis is to check whether conserved residues line up in

multiple structural alignments and how conserved residues and ligands cluster

together in multiple structure superimpositions. In favourable cases, protein

structure comparison can lead to evolutionary discoveries not detected by

sequence analysis.

DOI: 10.1093/nar/gkq366

PMCID: PMC2896194

PMID: 20457744 [Indexed for MEDLINE]

1966. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W41-5. doi: 10.1093/nar/gkq293.

Epub 2010 May 5.

ALADYN: a web server for aligning proteins by matching their large-scale motion.

Potestio R(1), Aleksiev T, Pontiggia F, Cozzini S, Micheletti C.

Author information:

(1)Scuola Internazionale Superiore di Studi Avanzati and eLab, via Bonomea 265,

34136 Trieste, Italy.

The ALADYN web server aligns pairs of protein structures by comparing their

internal dynamics and detecting regions that sustain similar large-scale

movements. The latter often accompany functional conformational changes in

proteins and enzymes. The ALADYN dynamics-based alignment can therefore highlight

functionally-oriented correspondences that could be more elusive to sequence- or

structure-based comparisons. The ALADYN server takes the structure files of the

two proteins as input. The optimal relative positioning of the molecules is found

by maximizing the similarity of the pattern of structural fluctuations which are

calculated via an elastic network model. The resulting alignment is presented via

an interactive graphical Java applet and is accompanied by a number of

quantitative indicators and downloadable data files. The ALADYN web server is

freely accessible at the http://aladyn.escience-lab.org address.

DOI: 10.1093/nar/gkq293

PMCID: PMC2896196

PMID: 20444876 [Indexed for MEDLINE]

1967. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W373-7. doi: 10.1093/nar/gkq316.

Epub 2010 May 5.

Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA.

Smith C(1), Heyne S, Richter AS, Will S, Backofen R.

Author information:

(1)Bioinformatics Group, University of Freiburg, Freiburg 79110, Germany.

The Freiburg RNA tools web server integrates three tools for the advanced

analysis of RNA in a common web-based user interface. The tools IntaRNA, ExpaRNA

and LocARNA support the prediction of RNA-RNA interaction, exact RNA matching and

alignment of RNA, respectively. The Freiburg RNA tools web server and the

software packages of the stand-alone tools are freely accessible at

http://rna.informatik.uni-freiburg.de.

DOI: 10.1093/nar/gkq316

PMCID: PMC2896085

PMID: 20444875 [Indexed for MEDLINE]

1968. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W622-7. doi: 10.1093/nar/gkq325.

Epub 2010 May 5.

Frog2: Efficient 3D conformation ensemble generator for small compounds.

Miteva MA(1), Guyon F, Tufféry P.

Author information:

(1)MTi, INSERM UMR-S973, Université Paris Diderot-Paris 7, Bat. Lamarck case

7113, 35 rue H. Brion, F75205, Paris, France.

Frog is a web tool dedicated to small compound 3D generation. Here we present the

new version, Frog2, which allows the generation of conformation ensembles of

small molecules starting from either 1D, 2D or 3D description of the compounds.

From a compound description in one of the SMILES, SDF or mol2 formats, the server

will return an ensemble of diverse conformers generated using a two stage Monte

Carlo approach in the dihedral space. When starting from 1D or 2D description of

compounds, Frog2 is capable to detect the sites of ambiguous stereoisomery, and

thus to sample different stereoisomers. Frog2 also embeds new energy minimization

and ring generation facilities that solve the problem of some missing cycle

structures in the Frog1 ring library. Finally, the optimized generator of

conformation ensembles in Frog2 results in a gain of computational time

permitting Frog2 to be up to 20 times faster that Frog1, while producing

satisfactory conformations in terms of structural quality and conformational

diversity. The high speed and the good quality of generated conformational

ensembles makes it possible the treatment of larger compound collections using

Frog2. The server and documentation are freely available at

http://bioserv.rpbs.univ-paris-diderot.fr/Frog2.

DOI: 10.1093/nar/gkq325

PMCID: PMC2896087

PMID: 20444874 [Indexed for MEDLINE]

1969. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W84-9. doi: 10.1093/nar/gkq320.

Epub 2010 May 5.

PANDORA: analysis of protein and peptide sets through the hierarchical

integration of annotations.

Rappoport N(1), Fromer M, Schweiger R, Linial M.

Author information:

(1)School of Computer Science and Engineering, The Hebrew University of

Jerusalem, Israel.

Derivation of biological meaning from large sets of proteins or genes is a

frequent task in genomic and proteomic studies. Such sets often arise from

experimental methods including large-scale gene expression experiments and mass

spectrometry (MS) proteomics. Large sets of genes or proteins are also the

outcome of computational methods such as BLAST search and homology-based

classifications. We have developed the PANDORA web server, which functions as a

platform for the advanced biological analysis of sets of genes, proteins, or

proteolytic peptides. First, the input set is mapped to a set of corresponding

proteins. Then, an analysis of the protein set produces a graph-based hierarchy

which highlights intrinsic relations amongst biological subsets, in light of

their different annotations from multiple annotation resources. PANDORA

integrates a large collection of annotation sources (GO, UniProt Keywords,

InterPro, Enzyme, SCOP, CATH, Gene-3D, NCBI taxonomy and more) that comprise

approximately 200,000 different annotation terms associated with approximately

3.2 million sequences from UniProtKB. Statistical enrichment based on a binomial

approximation of the hypergeometric distribution and corrected for multiple

hypothesis tests is calculated using several background sets, including major

gene-expression DNA-chip platforms. Users can also visualize either standard or

user-defined binary and quantitative properties alongside the proteins. PANDORA

4.2 is available at http://www.pandora.cs.huji.ac.il.

DOI: 10.1093/nar/gkq320

PMCID: PMC2896089

PMID: 20444873 [Indexed for MEDLINE]

1970. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W402-6. doi: 10.1093/nar/gkq323.

Epub 2010 May 5.

HotPoint: hot spot prediction server for protein interfaces.

Tuncbag N(1), Keskin O, Gursoy A.

Author information:

(1)Center for Computational Biology and Bioinformatics, and College of

Engineering, Koc University, Rumelifeneri Yolu, 34450 Sariyer Istanbul, Turkey.

The energy distribution along the protein-protein interface is not homogenous;

certain residues contribute more to the binding free energy, called 'hot spots'.

Here, we present a web server, HotPoint, which predicts hot spots in protein

interfaces using an empirical model. The empirical model incorporates a few

simple rules consisting of occlusion from solvent and total knowledge-based pair

potentials of residues. The prediction model is computationally efficient and

achieves high accuracy of 70%. The input to the HotPoint server is a protein

complex and two chain identifiers that form an interface. The server provides the

hot spot prediction results, a table of residue properties and an interactive 3D

visualization of the complex with hot spots highlighted. Results are also

downloadable as text files. This web server can be used for analysis of any

protein-protein interface which can be utilized by researchers working on binding

sites characterization and rational design of small molecules for protein

interactions. HotPoint is accessible at http://prism.ccbb.ku.edu.tr/hotpoint.

DOI: 10.1093/nar/gkq323

PMCID: PMC2896123

PMID: 20444871 [Indexed for MEDLINE]

1971. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W445-9. doi: 10.1093/nar/gkq311.

Epub 2010 May 5.

HexServer: an FFT-based protein docking server powered by graphics processors.

Macindoe G(1), Mavridis L, Venkatraman V, Devignes MD, Ritchie DW.

Author information:

(1)Department of Computing Science, Lillybank Gardens, University of Glasgow, G12

8QQ Scotland, UK.

HexServer (http://hexserver.loria.fr/) is the first Fourier transform (FFT)-based

protein docking server to be powered by graphics processors. Using two graphics

processors simultaneously, a typical 6D docking run takes approximately 15 s,

which is up to two orders of magnitude faster than conventional FFT-based docking

approaches using comparable resolution and scoring functions. The server requires

two protein structures in PDB format to be uploaded, and it produces a ranked

list of up to 1000 docking predictions. Knowledge of one or both protein binding

sites may be used to focus and shorten the calculation when such information is

available. The first 20 predictions may be accessed individually, and a single

file of all predicted orientations may be downloaded as a compressed multi-model

PDB file. The server is publicly available and does not require any registration

or identification by the user.

DOI: 10.1093/nar/gkq311

PMCID: PMC2896144

PMID: 20444869 [Indexed for MEDLINE]

1972. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W615-21. doi:

10.1093/nar/gkq322. Epub 2010 May 5.

e-LEA3D: a computational-aided drug design web server.

Douguet D(1).

Author information:

(1)CNRS UMR6097-Université Nice-Sophia Antipolis 660, route des lucioles 06560

Valbonne, France. douguet@ipmc.cnrs.fr

e-LEA3D web server integrates three complementary tools to perform computer-aided

drug design based on molecular fragments. In drug discovery projects, there is a

considerable interest in identifying novel and diverse molecular scaffolds to

enhance chances of success. The de novo drug design tool is used to invent new

ligands to optimize a user-specified scoring function. The composite scoring

function includes both structure- and ligand-based evaluations. The de novo

approach is an alternative to a blind virtual screening of large compound

collections. A heuristic based on a genetic algorithm rapidly finds which

fragments or combination of fragments fit a QSAR model or the binding site of a

protein. While the approach is ideally suited for scaffold-hopping, this module

also allows a scan for possible substituents to a user-specified scaffold. The

second tool offers a traditional virtual screening and filtering of an uploaded

library of compounds. The third module addresses the combinatorial library design

that is based on a user-drawn scaffold and reactants coming, for example, from a

chemical supplier. The e-LEA3D server is available at:

http://bioinfo.ipmc.cnrs.fr/lea.html.

DOI: 10.1093/nar/gkq322

PMCID: PMC2896156

PMID: 20444867 [Indexed for MEDLINE]

1973. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W132-7. doi: 10.1093/nar/gkq312.

Epub 2010 May 5.

MetExplore: a web server to link metabolomic experiments and genome-scale

metabolic networks.

Cottret L(1), Wildridge D, Vinson F, Barrett MP, Charles H, Sagot MF, Jourdan F.

Author information:

(1)INRA, UMR1089, Xénobiotiques, F-31000 Toulouse, France.

ludovic.cottret@toulouse.inra.fr

High-throughput metabolomic experiments aim at identifying and ultimately

quantifying all metabolites present in biological systems. The metabolites are

interconnected through metabolic reactions, generally grouped into metabolic

pathways. Classical metabolic maps provide a relational context to help interpret

metabolomics experiments and a wide range of tools have been developed to help

place metabolites within metabolic pathways. However, the representation of

metabolites within separate disconnected pathways overlooks most of the

connectivity of the metabolome. By definition, reference pathways cannot

integrate novel pathways nor show relationships between metabolites that may be

linked by common neighbours without being considered as joint members of a

classical biochemical pathway. MetExplore is a web server that offers the

possibility to link metabolites identified in untargeted metabolomics experiments

within the context of genome-scale reconstructed metabolic networks. The analysis

pipeline comprises mapping metabolomics data onto the specific metabolic network

of an organism, then applying graph-based methods and advanced visualization

tools to enhance data analysis. The MetExplore web server is freely accessible at

http://metexplore.toulouse.inra.fr.

DOI: 10.1093/nar/gkq312

PMCID: PMC2896158

PMID: 20444866 [Indexed for MEDLINE]

1974. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W450-6. doi: 10.1093/nar/gkq328.

Epub 2010 Apr 30.

ConPlex: a server for the evolutionary conservation analysis of protein complex

structures.

Choi YS(1), Han SK, Kim J, Yang JS, Jeon J, Ryu SH, Kim S.

Author information:

(1)School of Interdisciplinary Bioscience and Bioengineering, Pohang University

of Science and Technology, Pohang, 790-784, Republic of Korea.

Evolutionary conservation analyses are important for the identification of

protein-protein interactions. For protein complex structures, sequence

conservation has been applied to determine protein oligomerization states, to

characterize native interfaces from non-specific crystal contacts, and to

discriminate near-native structures from docking artifacts. However, a

user-friendly web-based service for evolutionary conservation analysis of protein

complexes has not been available. Therefore, we developed ConPlex

(http://sbi.postech.ac.kr/ConPlex/) a web application that enables evolutionary

conservation analyses of protein interactions within protein quaternary

structures. Users provide protein complex structures; ConPlex automatically

identifies protein interfaces and carries out evolutionary conservation analyses

for the interface regions. Moreover, ConPlex allows the results of the

residue-specific conservation analysis to be displayed on the protein complex

structure and provides several options to customize the display output to fit

each user's needs. We believe that ConPlex offers a convenient platform to

analyze protein complex structures based on evolutionary conservation of

protein-protein interface residues.

DOI: 10.1093/nar/gkq328

PMCID: PMC2896159

PMID: 20435678 [Indexed for MEDLINE]

1975. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W7-13. doi: 10.1093/nar/gkq291.

Epub 2010 Apr 30.

TranslatorX: multiple alignment of nucleotide sequences guided by amino acid

translations.

Abascal F(1), Zardoya R, Telford MJ.

Author information:

(1)Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias

Naturales, CSIC, José Gutiérrez Abascal, 2, 28006 Madrid, Spain.

We present TranslatorX, a web server designed to align protein-coding nucleotide

sequences based on their corresponding amino acid translations. Many comparisons

between biological sequences (nucleic acids and proteins) involve the

construction of multiple alignments. Alignments represent a statement regarding

the homology between individual nucleotides or amino acids within homologous

genes. As protein-coding DNA sequences evolve as triplets of nucleotides (codons)

and it is known that sequence similarity degrades more rapidly at the DNA than at

the amino acid level, alignments are generally more accurate when based on amino

acids than on their corresponding nucleotides. TranslatorX novelties include: (i)

use of all documented genetic codes and the possibility of assigning different

genetic codes for each sequence; (ii) a battery of different multiple alignment

programs; (iii) translation of ambiguous codons when possible; (iv) an innovative

criterion to clean nucleotide alignments with GBlocks based on protein

information; and (v) a rich output, including Jalview-powered graphical

visualization of the alignments, codon-based alignments coloured according to the

corresponding amino acids, measures of compositional bias and first, second and

third codon position specific alignments. The TranslatorX server is freely

available at http://translatorx.co.uk.

DOI: 10.1093/nar/gkq291

PMCID: PMC2896173

PMID: 20435676 [Indexed for MEDLINE]

1976. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W90-5. doi: 10.1093/nar/gkq324.

Epub 2010 Apr 30.

i-GSEA4GWAS: a web server for identification of pathways/gene sets associated

with traits by applying an improved gene set enrichment analysis to genome-wide

association study.

Zhang K(1), Cui S, Chang S, Zhang L, Wang J.

Author information:

(1)Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of

Sciences, 100101, Beijing, China.

Genome-wide association study (GWAS) is nowadays widely used to identify genes

involved in human complex disease. The standard GWAS analysis examines SNPs/genes

independently and identifies only a number of the most significant SNPs. It

ignores the combined effect of weaker SNPs/genes, which leads to difficulties to

explore biological function and mechanism from a systems point of view. Although

gene set enrichment analysis (GSEA) has been introduced to GWAS to overcome these

limitations by identifying the correlation between pathways/gene sets and traits,

the heavy dependence on genotype data, which is not easily available for most

published GWAS investigations, has led to limited application of it. In order to

perform GSEA on a simple list of GWAS SNP P-values, we implemented GSEA by using

SNP label permutation. We further improved GSEA (i-GSEA) by focusing on

pathways/gene sets with high proportion of significant genes. To provide

researchers an open platform to analyze GWAS data, we developed the i-GSEA4GWAS

(improved GSEA for GWAS) web server. i-GSEA4GWAS implements the i-GSEA approach

and aims to provide new insights in complex disease studies. i-GSEA4GWAS is

freely available at http://gsea4gwas.psych.ac.cn/.

DOI: 10.1093/nar/gkq324

PMCID: PMC2896119

PMID: 20435672 [Indexed for MEDLINE]

1977. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W138-43. doi:

10.1093/nar/gkq318. Epub 2010 Apr 30.

PathPred: an enzyme-catalyzed metabolic pathway prediction server.

Moriya Y(1), Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, Kanehisa M.

Author information:

(1)Bioinformatics Center, Institute for Chemical Research, Kyoto University,

Gokasho, Uji, Kyoto 611-0011, Japan.

The KEGG RPAIR database is a collection of biochemical structure transformation

patterns, called RDM patterns, and chemical structure alignments of

substrate-product pairs (reactant pairs) in all known enzyme-catalyzed reactions

taken from the Enzyme Nomenclature and the KEGG PATHWAY database. Here, we

present PathPred (http://www.genome.jp/tools/pathpred/), a web-based server to

predict plausible pathways of muti-step reactions starting from a query compound,

based on the local RDM pattern match and the global chemical structure alignment

against the reactant pair library. In this server, we focus on predicting

pathways for microbial biodegradation of environmental compounds and biosynthesis

of plant secondary metabolites, which correspond to characteristic RDM patterns

in 947 and 1397 reactant pairs, respectively. The server provides transformed

compounds and reference transformation patterns in each predicted reaction, and

displays all predicted multi-step reaction pathways in a tree-shaped graph.

DOI: 10.1093/nar/gkq318

PMCID: PMC2896155

PMID: 20435670 [Indexed for MEDLINE]

1978. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W609-14. doi:

10.1093/nar/gkq300. Epub 2010 Apr 29.

PharmMapper server: a web server for potential drug target identification using

pharmacophore mapping approach.

Liu X(1), Ouyang S, Yu B, Liu Y, Huang K, Gong J, Zheng S, Li Z, Li H, Jiang H.

Author information:

(1)Drug Discovery and Design Center, Shanghai Institute of Materia Medica,

Chinese Academy of Sciences, Shanghai 201203, China.

In silico drug target identification, which includes many distinct algorithms for

finding disease genes and proteins, is the first step in the drug discovery

pipeline. When the 3D structures of the targets are available, the problem of

target identification is usually converted to finding the best interaction mode

between the potential target candidates and small molecule probes. Pharmacophore,

which is the spatial arrangement of features essential for a molecule to interact

with a specific target receptor, is an alternative method for achieving this goal

apart from molecular docking method. PharmMapper server is a freely accessed web

server designed to identify potential target candidates for the given small

molecules (drugs, natural products or other newly discovered compounds with

unidentified binding targets) using pharmacophore mapping approach. PharmMapper

hosts a large, in-house repertoire of pharmacophore database (namely

PharmTargetDB) annotated from all the targets information in TargetBank,

BindingDB, DrugBank and potential drug target database, including over 7000

receptor-based pharmacophore models (covering over 1500 drug targets

information). PharmMapper automatically finds the best mapping poses of the query

molecule against all the pharmacophore models in PharmTargetDB and lists the top

N best-fitted hits with appropriate target annotations, as well as respective

molecule's aligned poses are presented. Benefited from the highly efficient and

robust triangle hashing mapping method, PharmMapper bears high throughput ability

and only costs 1 h averagely to screen the whole PharmTargetDB. The protocol was

successful in finding the proper targets among the top 300 pharmacophore

candidates in the retrospective benchmarking test of tamoxifen. PharmMapper is

available at http://59.78.96.61/pharmmapper.

DOI: 10.1093/nar/gkq300

PMCID: PMC2896160

PMID: 20430828 [Indexed for MEDLINE]

1979. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W29-34. doi: 10.1093/nar/gkq298.

Epub 2010 Apr 29.

SATCHMO-JS: a webserver for simultaneous protein multiple sequence alignment and

phylogenetic tree construction.

Hagopian R(1), Davidson JR, Datta RS, Samad B, Jarvis GR, Sjölander K.

Author information:

(1)Department of Bioengineering, University of California, Berkeley, CA 94720,

USA.

We present the jump-start simultaneous alignment and tree construction using

hidden Markov models (SATCHMO-JS) web server for simultaneous estimation of

protein multiple sequence alignments (MSAs) and phylogenetic trees. The server

takes as input a set of sequences in FASTA format, and outputs a phylogenetic

tree and MSA; these can be viewed online or downloaded from the website.

SATCHMO-JS is an extension of the SATCHMO algorithm, and employs a

divide-and-conquer strategy to jump-start SATCHMO at a higher point in the

phylogenetic tree, reducing the computational complexity of the progressive

all-versus-all HMM-HMM scoring and alignment. Results on a benchmark dataset of

983 structurally aligned pairs from the PREFAB benchmark dataset show that

SATCHMO-JS provides a statistically significant improvement in alignment accuracy

over MUSCLE, Multiple Alignment using Fast Fourier Transform (MAFFT), ClustalW

and the original SATCHMO algorithm. The SATCHMO-JS webserver is available at

http://phylogenomics.berkeley.edu/satchmo-js. The datasets used in these

experiments are available for download at

http://phylogenomics.berkeley.edu/satchmo-js/supplementary/.

DOI: 10.1093/nar/gkq298

PMCID: PMC2896197

PMID: 20430824 [Indexed for MEDLINE]

1980. PLoS One. 2010 Jun 28;5(6):e11335. doi: 10.1371/journal.pone.0011335.

Plant-mPLoc: a top-down strategy to augment the power for predicting plant

protein subcellular localization.

Chou KC(1), Shen HB.

Author information:

(1)Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong

University, Shanghai, China. kcchou@gordonlifescience.org

One of the fundamental goals in proteomics and cell biology is to identify the

functions of proteins in various cellular organelles and pathways. Information of

subcellular locations of proteins can provide useful insights for revealing their

functions and understanding how they interact with each other in cellular network

systems. Most of the existing methods in predicting plant protein subcellular

localization can only cover three or four location sites, and none of them can be

used to deal with multiplex plant proteins that can simultaneously exist at two,

or move between, two or more different location sits. Actually, such multiplex

proteins might have special biological functions worthy of particular notice. The

present study was devoted to improve the existing plant protein subcellular

location predictors from the aforementioned two aspects. A new predictor called

"Plant-mPLoc" is developed by integrating the gene ontology information,

functional domain information, and sequential evolutionary information through

three different modes of pseudo amino acid composition. It can be used to

identify plant proteins among the following 12 location sites: (1) cell membrane,

(2) cell wall, (3) chloroplast, (4) cytoplasm, (5) endoplasmic reticulum, (6)

extracellular, (7) Golgi apparatus, (8) mitochondrion, (9) nucleus, (10)

peroxisome, (11) plastid, and (12) vacuole. Compared with the existing methods

for predicting plant protein subcellular localization, the new predictor is much

more powerful and flexible. Particularly, it also has the capacity to deal with

multiple-location proteins, which is beyond the reach of any existing predictors

specialized for identifying plant protein subcellular localization. As a

user-friendly web-server, Plant-mPLoc is freely accessible at

http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/. Moreover, for the convenience

of the vast majority of experimental scientists, a step-by-step guide is provided

on how to use the web-server to get the desired results. It is anticipated that

the Plant-mPLoc predictor as presented in this paper will become a very useful

tool in plant science as well as all the relevant areas.

DOI: 10.1371/journal.pone.0011335

PMCID: PMC2893129

PMID: 20596258 [Indexed for MEDLINE]

1981. Biochem Biophys Res Commun. 2010 Jun 25;397(2):340-4. doi:

10.1016/j.bbrc.2010.05.125. Epub 2010 May 27.

HapAssembler: a web server for haplotype assembly from SNP fragments using

genetic algorithm.

Kang SH(1), Jeong IS, Cho HG, Lim HS.

Author information:

(1)School of Electronics & Computer Engineering, Chonnam National University, 77

Yongbong-ro, Buk-gu, Gwangju 500-757, Republic of Korea.

Haplotype, which is the sequence of SNPs in a specific chromosome, plays an

important role in disease association studies. However, current sequencing

techniques can detect the presence of SNP sites, but they cannot tell which copy

of a pair of chromosomes the alleles belong to. Moreover, sequencing errors that

occurred in sequencing SNP fragments make it difficult to determine a pair of

haplotypes from SNP fragments. To help overcome this difficulty, the haplotype

assembly problem is defined from the viewpoint of computation, and several models

are suggested to tackle this problem. However, there are no freely available

web-based tools to overcome this problem as far as we are aware. In this paper,

we present a web-based application based on the genetic algorithm, named

HapAssembler, for assembling a pair of haplotypes from SNP fragments. Numerical

results on real biological data show that the correct rate of the proposed

application in this paper is greater than 95% in most cases. HapAssembler is

freely available at http://alex.chonnam.ac.kr/~drminor/hapHome.htm. Users can

choose any model among four models for their purpose and determine haplotypes

from their input data.

Copyright (c) 2010 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.bbrc.2010.05.125

PMID: 20510878 [Indexed for MEDLINE]

1982. BMC Bioinformatics. 2010 Jun 18;11:333. doi: 10.1186/1471-2105-11-333.

MPRAP: an accessibility predictor for a-helical transmembrane proteins that

performs well inside and outside the membrane.

Illergård K(1), Callegari S, Elofsson A.

Author information:

(1)Center for Biomembrane Research, Stockholm Bioinformatics Center, Dept of

Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden.

BACKGROUND: In water-soluble proteins it is energetically favorable to bury

hydrophobic residues and to expose polar and charged residues. In contrast to

water soluble proteins, transmembrane proteins face three distinct environments;

a hydrophobic lipid environment inside the membrane, a hydrophilic water

environment outside the membrane and an interface region rich in phospholipid

head-groups. Therefore, it is energetically favorable for transmembrane proteins

to expose different types of residues in the different regions.

RESULTS: Investigations of a set of structurally determined transmembrane

proteins showed that the composition of solvent exposed residues differs

significantly inside and outside the membrane. In contrast, residues buried

within the interior of a protein show a much smaller difference. However, in all

regions exposed residues are less conserved than buried residues. Further, we

found that current state-of-the-art predictors for surface area are optimized for

one of the regions and perform badly in the other regions. To circumvent this

limitation we developed a new predictor, MPRAP, that performs well in all

regions. In addition, MPRAP performs better on complete membrane proteins than a

combination of specialized predictors and acceptably on water-soluble proteins. A

web-server of MPRAP is available at http://mprap.cbr.su.se/

CONCLUSION: By including complete a-helical transmembrane proteins in the

training MPRAP is able to predict surface accessibility accurately both inside

and outside the membrane. This predictor can aid in the prediction of

3D-structure, and in the identification of erroneous protein structures.

DOI: 10.1186/1471-2105-11-333

PMCID: PMC2904353

PMID: 20565847 [Indexed for MEDLINE]

1983. Bioinformatics. 2010 Jun 15;26(12):i278-86. doi: 10.1093/bioinformatics/btq218.

Thermodynamics of RNA structures by Wang-Landau sampling.

Lou F(1), Clote P.

Author information:

(1)Laboratoire de Recherche en Informatique, Université Paris-Sud XI, bât. 490,

91405 Orsay cedex, France.

MOTIVATION: Thermodynamics-based dynamic programming RNA secondary structure

algorithms have been of immense importance in molecular biology, where

applications range from the detection of novel selenoproteins using expressed

sequence tag (EST) data, to the determination of microRNA genes and their

targets. Dynamic programming algorithms have been developed to compute the

minimum free energy secondary structure and partition function of a given RNA

sequence, the minimum free-energy and partition function for the hybridization of

two RNA molecules, etc. However, the applicability of dynamic programming methods

depends on disallowing certain types of interactions (pseudoknots, zig-zags,

etc.), as their inclusion renders structure prediction an nondeterministic

polynomial time (NP)-complete problem. Nevertheless, such interactions have been

observed in X-ray structures.

RESULTS: A non-Boltzmannian Monte Carlo algorithm was designed by Wang and Landau

to estimate the density of states for complex systems, such as the Ising model,

that exhibit a phase transition. In this article, we apply the Wang-Landau (WL)

method to compute the density of states for secondary structures of a given RNA

sequence, and for hybridizations of two RNA sequences. Our method is shown to be

much faster than existent software, such as RNAsubopt. From density of states, we

compute the partition function over all secondary structures and over all

pseudoknot-free hybridizations. The advantage of the WL method is that by adding

a function to evaluate the free energy of arbitrary pseudoknotted structures and

of arbitrary hybridizations, we can estimate thermodynamic parameters for

situations known to be NP-complete. This extension to pseudoknots will be made in

the sequel to this article; in contrast, the current article describes the WL

algorithm applied to pseudoknot-free secondary structures and hybridizations.

AVAILABILITY: The WL RNA hybridization web server is under construction at

http://bioinformatics.bc.edu/clotelab/.

DOI: 10.1093/bioinformatics/btq218

PMCID: PMC2881402

PMID: 20529917 [Indexed for MEDLINE]

1984. Bioinformatics. 2010 Jun 15;26(12):1566-8. doi: 10.1093/bioinformatics/btq233.

Epub 2010 Apr 28.

TAPIR, a web server for the prediction of plant microRNA targets, including

target mimics.

Bonnet E(1), He Y, Billiau K, Van de Peer Y.

Author information:

(1)Department of Plant Systems Biology, VIB, B-9052 Gent, Belgium.

We present a new web server called TAPIR, designed for the prediction of plant

microRNA targets. The server offers the possibility to search for plant miRNA

targets using a fast and a precise algorithm. The precise option is much slower

but guarantees to find less perfectly paired miRNA-target duplexes. Furthermore,

the precise option allows the prediction of target mimics, which are

characterized by a miRNA-target duplex having a large loop, making them

undetectable by traditional tools.AVAILABILITY: The TAPIR web server can be

accessed at: http://bioinformatics.psb.ugent.be/webtools/tapir.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq233

PMID: 20430753 [Indexed for MEDLINE]

1985. Bioinformatics. 2010 Jun 15;26(12):1528-34. doi: 10.1093/bioinformatics/btq141.

Epub 2010 Apr 12.

Modular rate laws for enzymatic reactions: thermodynamics, elasticities and

implementation.

Liebermeister W(1), Uhlendorf J, Klipp E.

Author information:

(1)Institut für Biologie, Theoretische Biophysik, Humboldt-Universität zu Berlin,

Berlin, Germany. wolfram.liebermeister@biologie.hu-berlin.de

MOTIVATION: Standard rate laws are a key requisite for systematically turning

metabolic networks into kinetic models. They should provide simple, general and

biochemically plausible formulae for reaction velocities and reaction

elasticities. At the same time, they need to respect thermodynamic relations

between the kinetic constants and the metabolic fluxes and concentrations.

RESULTS: We present a family of reversible rate laws for reactions with arbitrary

stoichiometries and various types of regulation, including mass-action,

Michaelis-Menten and uni-uni reversible Hill kinetics as special cases. With a

thermodynamically safe parameterization of these rate laws, parameter sets

obtained by model fitting, sampling or optimization are guaranteed to lead to

consistent chemical equilibrium states. A reformulation using saturation values

yields simple formulae for rates and elasticities, which can be easily adjusted

to the given stationary flux distributions. Furthermore, this formulation

highlights the role of chemical potential differences as thermodynamic driving

forces. We compare the modular rate laws to the thermodynamic-kinetic modelling

formalism and discuss a simplified rate law in which the reaction rate directly

depends on the reaction affinity. For automatic handling of modular rate laws, we

propose a standard syntax and semantic annotations for the Systems Biology Markup

Language.

AVAILABILITY: An online tool for inserting the rate laws into SBML models is

freely available at www.semanticsbml.org.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq141

PMID: 20385728 [Indexed for MEDLINE]

1986. BMC Bioinformatics. 2010 Jun 14;11:318. doi: 10.1186/1471-2105-11-318.

CONAN: copy number variation analysis software for genome-wide association

studies.

Forer L(1), Schönherr S, Weissensteiner H, Haider F, Kluckner T, Gieger C,

Wichmann HE, Specht G, Kronenberg F, Kloss-Brandstätter A.

Author information:

(1)Division of Genetic Epidemiology, Department of Medical Genetics, Molecular

and Clinical Pharmacology, Innsbruck Medical University, 6020 Innsbruck, Austria.

BACKGROUND: Genome-wide association studies (GWAS) based on single nucleotide

polymorphisms (SNPs) revolutionized our perception of the genetic regulation of

complex traits and diseases. Copy number variations (CNVs) promise to shed

additional light on the genetic basis of monogenic as well as complex diseases

and phenotypes. Indeed, the number of detected associations between CNVs and

certain phenotypes are constantly increasing. However, while several software

packages support the determination of CNVs from SNP chip data, the downstream

statistical inference of CNV-phenotype associations is still subject to

complicated and inefficient in-house solutions, thus strongly limiting the

performance of GWAS based on CNVs.

RESULTS: CONAN is a freely available client-server software solution which

provides an intuitive graphical user interface for categorizing, analyzing and

associating CNVs with phenotypes. Moreover, CONAN assists the evaluation process

by visualizing detected associations via Manhattan plots in order to enable a

rapid identification of genome-wide significant CNV regions. Various file formats

including the information on CNVs in population samples are supported as input

data.

CONCLUSIONS: CONAN facilitates the performance of GWAS based on CNVs and the

visual analysis of calculated results. CONAN provides a rapid, valid and

straightforward software solution to identify genetic variation underlying the

'missing' heritability for complex traits that remains unexplained by recent

GWAS. The freely available software can be downloaded at

http://genepi-conan.i-med.ac.at.

DOI: 10.1186/1471-2105-11-318

PMCID: PMC2894823

PMID: 20546565 [Indexed for MEDLINE]

1987. BMC Bioinformatics. 2010 Jun 10;11:315. doi: 10.1186/1471-2105-11-315.

mu-CS: an extension of the TM4 platform to manage Affymetrix binary data.

Guzzi PH(1), Cannataro M.

Author information:

(1)Bioinformatics Laboratory, Department of Experimental Medicine and Clinic,

Magna Graecia University, Catanzaro, Italy. hguzzi@unicz.it

BACKGROUND: A main goal in understanding cell mechanisms is to explain the

relationship among genes and related molecular processes through the combined use

of technological platforms and bioinformatics analysis. High throughput

platforms, such as microarrays, enable the investigation of the whole genome in a

single experiment. There exist different kind of microarray platforms, that

produce different types of binary data (images and raw data). Moreover, also

considering a single vendor, different chips are available. The analysis of

microarray data requires an initial preprocessing phase (i.e. normalization and

summarization) of raw data that makes them suitable for use on existing

platforms, such as the TIGR M4 Suite. Nevertheless, the annotations of data with

additional information such as gene function, is needed to perform more powerful

analysis. Raw data preprocessing and annotation is often performed in a manual

and error prone way. Moreover, many available preprocessing tools do not support

annotation. Thus novel, platform independent, and possibly open source tools

enabling the semi-automatic preprocessing and annotation of microarray data are

needed.

RESULTS: The paper presents mu-CS (Microarray Cel file Summarizer), a

cross-platform tool for the automatic normalization, summarization and annotation

of Affymetrix binary data. mu-CS is based on a client-server architecture. The

mu-CS client is provided both as a plug-in of the TIGR M4 platform and as a Java

standalone tool and enables users to read, preprocess and analyse binary

microarray data, avoiding the manual invocation of external tools (e.g. the

Affymetrix Power Tools), the manual loading of preprocessing libraries, and the

management of intermediate files. The mu-CS server automatically updates the

references to the summarization and annotation libraries that are provided to the

mu-CS client before the preprocessing. The mu-CS server is based on the web

services technology and can be easily extended to support more microarray vendors

(e.g. Illumina).

CONCLUSIONS: Thus mu-CS users can directly manage binary data without worrying

about locating and invoking the proper preprocessing tools and chip-specific

libraries. Moreover, users of the mu-CS plugin for TM4 can manage Affymetrix

binary files without using external tools, such as APT (Affymetrix Power Tools)

and related libraries. Consequently, mu-CS offers four main advantages: (i) it

avoids to waste time for searching the correct libraries, (ii) it reduces

possible errors in the preprocessing and further analysis phases, e.g. due to the

incorrect choice of parameters or the use of old libraries, (iii) it implements

the annotation of preprocessed data, and finally, (iv) it may enhance the quality

of further analysis since it provides the most updated annotation libraries. The

mu-CS client is freely available as a plugin of the TM4 platform as well as a

standalone application at the project web site

(http://bioingegneria.unicz.it/M-CS).

DOI: 10.1186/1471-2105-11-315

PMCID: PMC2907348

PMID: 20537149 [Indexed for MEDLINE]

1988. BMC Evol Biol. 2010 Jun 6;10:167. doi: 10.1186/1471-2148-10-167.

WAMI: a web server for the analysis of minisatellite maps.

Abouelhoda M(1), El-Kalioby M, Giegerich R.

Author information:

(1)Center for Informatics Sciences, Nile University, Giza, Egypt.

mabouelhoda@yahoo.com

BACKGROUND: Minisatellites are genomic loci composed of tandem arrays of short

repetitive DNA segments. A minisatellite map is a sequence of symbols that

represents the tandem repeat array such that the set of symbols is in one-to-one

correspondence with the set of distinct repeats. Due to variations in repeat type

and organization as well as copy number, the minisatellite maps have been widely

used in forensic and population studies. In either domain, researchers need to

compare the set of maps to each other, to build phylogenetic trees, to spot

structural variations, and to study duplication dynamics. Efficient algorithms

for these tasks are required to carry them out reliably and in reasonable time.

RESULTS: In this paper we present WAMI, a web-server for the analysis of

minisatellite maps. It performs the above mentioned computational tasks using

efficient algorithms that take the model of map evolution into account. The WAMI

interface is easy to use and the results of each analysis task are visualized.

CONCLUSIONS: To the best of our knowledge, WAMI is the first server providing all

these computational facilities to the minisatellite community. The WAMI

web-interface and the source code of the underlying programs are available at

http://www.nubios.nileu.edu.eg/tools/wami.

DOI: 10.1186/1471-2148-10-167

PMCID: PMC2897807

PMID: 20525398 [Indexed for MEDLINE]

1989. BioData Min. 2010 Jun 4;3(1):3. doi: 10.1186/1756-0381-3-3.

Applications and methods utilizing the Simple Semantic Web Architecture and

Protocol (SSWAP) for bioinformatics resource discovery and disparate data and

service integration.

Nelson RT(1), Avraham S, Shoemaker RC, May GD, Ware D, Gessler DD.

Author information:

(1)University of Arizona, 1657 E, Helen St,, Tucson, AZ 85721, USA.

dgessler@iplantcollaborative.org.

BACKGROUND: Scientific data integration and computational service discovery are

challenges for the bioinformatic community. This process is made more difficult

by the separate and independent construction of biological databases, which makes

the exchange of data between information resources difficult and labor intensive.

A recently described semantic web protocol, the Simple Semantic Web Architecture

and Protocol (SSWAP; pronounced "swap") offers the ability to describe data and

services in a semantically meaningful way. We report how three major information

resources (Gramene, SoyBase and the Legume Information System [LIS]) used SSWAP

to semantically describe selected data and web services.

METHODS: We selected high-priority Quantitative Trait Locus (QTL), genomic

mapping, trait, phenotypic, and sequence data and associated services such as

BLAST for publication, data retrieval, and service invocation via semantic web

services. Data and services were mapped to concepts and categories as implemented

in legacy and de novo community ontologies. We used SSWAP to express these

offerings in OWL Web Ontology Language (OWL), Resource Description Framework

(RDF) and eXtensible Markup Language (XML) documents, which are appropriate for

their semantic discovery and retrieval. We implemented SSWAP services to respond

to web queries and return data. These services are registered with the SSWAP

Discovery Server and are available for semantic discovery at http://sswap.info.

RESULTS: A total of ten services delivering QTL information from Gramene were

created. From SoyBase, we created six services delivering information about

soybean QTLs, and seven services delivering genetic locus information. For LIS we

constructed three services, two of which allow the retrieval of DNA and RNA FASTA

sequences with the third service providing nucleic acid sequence comparison

capability (BLAST).

CONCLUSIONS: The need for semantic integration technologies has preceded

available solutions. We report the feasibility of mapping high priority data from

local, independent, idiosyncratic data schemas to common shared concepts as

implemented in web-accessible ontologies. These mappings are then amenable for

use in semantic web services. Our implementation of approximately two dozen

services means that biological data at three large information resources

(Gramene, SoyBase, and LIS) is available for programmatic access, semantic

searching, and enhanced interaction between the separate missions of these

resources.

DOI: 10.1186/1756-0381-3-3

PMCID: PMC2894815

PMID: 20525377

1990. BMC Bioinformatics. 2010 Jun 3;11:301. doi: 10.1186/1471-2105-11-301.

Prediction of GTP interacting residues, dipeptides and tripeptides in a protein

from its evolutionary information.

Chauhan JS(1), Mishra NK, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology (IMTECH), Chandigarh,

India.

BACKGROUND: Guanosine triphosphate (GTP)-binding proteins play an important role

in regulation of G-protein. Thus prediction of GTP interacting residues in a

protein is one of the major challenges in the field of the computational biology.

In this study, an attempt has been made to develop a computational method for

predicting GTP interacting residues in a protein with high accuracy (Acc),

precision (Prec) and recall (Rc).

RESULT: All the models developed in this study have been trained and tested on a

non-redundant (40% similarity) dataset using five-fold cross-validation. Firstly,

we have developed neural network based models using single sequence and PSSM

profile and achieved maximum Matthews Correlation Coefficient (MCC) 0.24 (Acc

61.30%) and 0.39 (Acc 68.88%) respectively. Secondly, we have developed a support

vector machine (SVM) based models using single sequence and PSSM profile and

achieved maximum MCC 0.37 (Prec 0.73, Rc 0.57, Acc 67.98%) and 0.55 (Prec 0.80,

Rc 0.73, Acc 77.17%) respectively. In this work, we have introduced a new concept

of predicting GTP interacting dipeptide (two consecutive GTP interacting

residues) and tripeptide (three consecutive GTP interacting residues) for the

first time. We have developed SVM based model for predicting GTP interacting

dipeptides using PSSM profile and achieved MCC 0.64 with precision 0.87, recall

0.74 and accuracy 81.37%. Similarly, SVM based model have been developed for

predicting GTP interacting tripeptides using PSSM profile and achieved MCC 0.70

with precision 0.93, recall 0.73 and accuracy 83.98%.

CONCLUSION: These results show that PSSM based method performs better than single

sequence based method. The prediction models based on dipeptides or tripeptides

are more accurate than the traditional model based on single residue. A web

server "GTPBinder" http://www.imtech.res.in/raghava/gtpbinder/ based on above

models has been developed for predicting GTP interacting residues in a protein.

DOI: 10.1186/1471-2105-11-301

PMCID: PMC3098072

PMID: 20525281 [Indexed for MEDLINE]

1991. Acta Crystallogr B. 2010 Jun;66(Pt 3):315-22. doi: 10.1107/S0108768110009031.

Epub 2010 May 6.

Symmetry analysis of extinction rules in diffuse-scattering experiments.

Withers RL(1), Aroyo MI, Perez-Mato JM, Orobengoa D.

Author information:

(1)Research School of Chemistry, ANU, Canberra ACT 0200, Australia.

Structured diffuse-scattering intensities, whether of compositional or of pure

displacive origin, static or dynamic, contain important information about the

symmetry of the individual compositional and/or displacive modes responsible for

the observed intensities. However, the interpretation of the experimental data is

very often impeded by the lack of a symmetry-based approach to the analysis of

the structured diffuse-scattering distributions. Recently, we have demonstrated

the existence of systematic phonon selection rules for diffuse scattering that

depend on the symmetries of the mode and the scattering vector, and not on the

specific structure. Here, we show that such symmetry analysis can be successfully

extended and also applied to structure-dependent diffuse scattering associated

with 'disordered' materials: the combination of theoretically determined,

diffuse-scattering extinction conditions with the concept of non-characteristic

orbits proves to be very useful in the interpretation of the observed

diffuse-scattering extinctions. The utility of this approach is illustrated by

the analysis of diffuse-scattering data from ThAsSe, FeOF and FeF(2). The

essential part of the associated calculations are performed by the computer

programs NEUTRON (systematic phonon extinction rules in inelastic scattering) and

NONCHAR (non-characteristic orbits of space groups) that are available on the

Bilbao crystallographic server (http://www.cryst.ehu.es).

DOI: 10.1107/S0108768110009031

PMID: 20484802

1992. Amino Acids. 2010 Jun;39(1):101-10. doi: 10.1007/s00726-009-0381-1. Epub 2009 Nov

12.

Prediction of mitochondrial proteins of malaria parasite using split amino acid

composition and PSSM profile.

Verma R(1), Varshney GC, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A,

Chandigarh, India. ruchi@imtech.res.in

The rate of human death due to malaria is increasing day-by-day. Thus the malaria

causing parasite Plasmodium falciparum (PF) remains the cause of concern. With

the wealth of data now available, it is imperative to understand protein

localization in order to gain deeper insight into their functional roles. In this

manuscript, an attempt has been made to develop prediction method for the

localization of mitochondrial proteins. In this study, we describe a method for

predicting mitochondrial proteins of malaria parasite using machine-learning

technique. All models were trained and tested on 175 proteins (40 mitochondrial

and 135 non-mitochondrial proteins) and evaluated using five-fold cross

validation. We developed a Support Vector Machine (SVM) model for predicting

mitochondrial proteins of P. falciparum, using amino acids and dipeptides

composition and achieved maximum MCC 0.38 and 0.51, respectively. In this study,

split amino acid composition (SAAC) is used where composition of N-termini,

C-termini, and rest of protein is computed separately. The performance of SVM

model improved significantly from MCC 0.38 to 0.73 when SAAC instead of simple

amino acid composition was used as input. In addition, SVM model has been

developed using composition of PSSM profile with MCC 0.75 and accuracy 91.38%. We

achieved maximum MCC 0.81 with accuracy 92% using a hybrid model, which combines

PSSM profile and SAAC. When evaluated on an independent dataset our method

performs better than existing methods. A web server PFMpred has been developed

for predicting mitochondrial proteins of malaria parasites (

http://www.imtech.res.in/raghava/pfmpred/).

DOI: 10.1007/s00726-009-0381-1

PMID: 19908123 [Indexed for MEDLINE]

1993. Artif Intell Med. 2010 Jun;49(2):127-32. doi: 10.1016/j.artmed.2010.03.004. Epub

2010 Apr 15.

PMirP: a pre-microRNA prediction method based on structure-sequence hybrid

features.

Zhao D(1), Wang Y, Luo D, Shi X, Wang L, Xu D, Yu J, Liang Y.

Author information:

(1)College of Computer Science and Technology, Jilin University, Key Laboratory

of Symbol Computation and Knowledge Engineering of the Ministry of Education,

Changchun 130012, PR China.

OBJECTIVE: MicroRNA is a type of small non-coding RNAs, which usually has a

stem-loop structure. As an important stage of microRNA, the pre-microRNA is

transported from nuclear to cytoplasm by exportin5 and finally cleaved into

mature microRNA. Structure-sequence features and minimum of free energy of

secondary structure have been used for predicting pre-microRNA. Meanwhile, the

double helix structure with free nucleotides and base-pairing features is used to

identify pre-miRNA for the first time.

METHODS: We applied support vector machine for a novel hybrid coding scheme using

left-triplet method, the free nucleotides, the minimum of free energy of

secondary structure and base-pairings features. Data sets of human pre-microRNA,

other 11 species and the latest pre-microRNA sequences were used for testing.

RESULTS: In this study we developed an improved method for pre-microRNA

prediction using a combination of various features and a web server called PMirP.

The prediction specificity and sensitivity for real and pseudo human

pre-microRNAs are as high as 98.4% and 94.9%, respectively. The web server is

freely available to the public at http://ccst.jlu.edu.cn/ci/bioinformatics/MiRNA

(accessed: 26 February 2010).

CONCLUSIONS: Experimental results show that the proposed method improves the

prediction efficiency and accuracy over existing methods. In addition, the PMirP

has lower computational complexity and higher throughput prediction capacity than

Mipred web server.

Copyright 2010 Elsevier B.V. All rights reserved.

DOI: 10.1016/j.artmed.2010.03.004

PMID: 20399081 [Indexed for MEDLINE]

1994. Bioinformatics. 2010 Jun 1;26(11):1465-7. doi: 10.1093/bioinformatics/btq161.

Epub 2010 Apr 15.

ParaSAM: a parallelized version of the significance analysis of microarrays

algorithm.

Sharma A(1), Zhao J, Podolsky R, McIndoe RA.

Author information:

(1)Center for Biotechnology and Genomic Medicine, School of Medicine, Medical

College of Georgia, Augusta, GA 30912, USA.

MOTIVATION: Significance analysis of microarrays (SAM) is a widely used

permutation-based approach to identifying differentially expressed genes in

microarray datasets. While SAM is freely available as an Excel plug-in and as an

R-package, analyses are often limited for large datasets due to very high memory

requirements.

SUMMARY: We have developed a parallelized version of the SAM algorithm called

ParaSAM to overcome the memory limitations. This high performance multithreaded

application provides the scientific community with an easy and manageable

client-server Windows application with graphical user interface and does not

require programming experience to run. The parallel nature of the application

comes from the use of web services to perform the permutations. Our results

indicate that ParaSAM is not only faster than the serial version, but also can

analyze extremely large datasets that cannot be performed using existing

implementations.

AVAILABILITY: A web version open to the public is available at

http://bioanalysis.genomics.mcg.edu/parasam. For local installations, both the

windows and web implementations of ParaSAM are available for free at

http://www.amdcc.org/bioinformatics/software/parasam.aspx.

DOI: 10.1093/bioinformatics/btq161

PMCID: PMC2872005

PMID: 20400455 [Indexed for MEDLINE]

1995. Curr Protoc Bioinformatics. 2010 Jun;Chapter 10:Unit10.4. doi:

10.1002/0471250953.bi1004s30.

MultiPipMaker: a comparative alignment server for multiple DNA sequences.

Elnitski L(1), Burhans R, Riemer C, Hardison R, Miller W.

Author information:

(1)The Pennsylvania State University, University Park, Pennsylvania, USA.

The MultiPipMaker World Wide Web server (http://www.bx.psu.edu) provides a tool

for aligning multiple DNA sequences and visualizing regions of conservation among

them. This unit describes its use and gives an explanation of the resulting

output files and supporting tools. Features provided by the server include

alignment of up to 20 very long genomic sequences, output choices of a true,

nucleotide-level multiple alignment and/or stacked, pairwise percent identity

plots, and support for user-specified annotations of genomic features and

arbitrary regions, with clickable links to additional information. Input

sequences other than the reference can be fragmented, unordered, and unoriented.

DOI: 10.1002/0471250953.bi1004s30

PMID: 20521245 [Indexed for MEDLINE]

1996. Curr Protoc Bioinformatics. 2010 Jun;Chapter 1:Unit1.15. doi:

10.1002/0471250953.bi0115s30.

Using the ensembl genome server to browse genomic sequence data.

Fernández-Suárez XM(1), Schuster MK.

Author information:

(1)EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton,

Cambridge, United Kingdom.

The Ensembl project provides a comprehensive source of automatic annotation of

the human genome sequence, as well as other species of biomedical interest, with

confirmed gene predictions that have been integrated with external data sources.

This unit describes how to use the Ensembl genome browser

(http://www.ensembl.org/), the public interface of the project. It describes how

to find a gene or protein of interest, how to get additional information and

external links, and how to use the comparative genomic data. Curr. Protoc.

Bioinform. 30:1.15.1-1.15.48. (c) 2010 by John Wiley & Sons, Inc.

DOI: 10.1002/0471250953.bi0115s30

PMID: 20521244 [Indexed for MEDLINE]

1997. Genomics Proteomics Bioinformatics. 2010 Jun;8(2):127-34. doi:

10.1016/S1672-0229(10)60014-9.

PsRNA: a computing engine for the comparative identification of putative small

RNA locations within intergenic regions.

Sridhar J(1), Sowmiya G, Sekar K, Rafi ZA.

Author information:

(1)Centre of Excellence in Bioinformatics, School of Biotechnology, Madurai

Kamaraj University, Madurai, India.

Small RNAs (sRNAs) are non-coding transcripts exerting their functions in the

cells directly. Identification of sRNAs is a difficult task due to the lack of

clear sequence and structural biases. Most sRNAs are identified within genus

specific intergenic regions in related genomes. However, several of these regions

remain un-annotated due to lack of sequence homology and/or potent statistical

identification tools. A computational engine has been built to search within the

intergenic regions to identify and roughly annotate new putative sRNA regions in

Enterobacteriaceae genomes. It utilizes experimentally known sRNA data and their

flanking genes/KEGG Orthology (KO) numbers as templates to identify similar sRNA

regions in related query genomes. The search engine not only has the capability

to locate putative intergenic regions for specific sRNAs, but also has the

potency to locate conserved, shuffled or deleted gene clusters in query genomes.

Because it uses the KO terms for locating functionally important regions such as

sRNAs, any further KO number assignment to additional genes will increase the

sensitivity. The PsRNA server is used for the identification of putative sRNA

regions through the information retrieved from the sRNA of interest. The

computing engine is available online at

http://bioserver1.physics.iisc.ernet.in/psrna/ and http://bicmku.in:8081/psrna/.

Copyright 2010 Beijing Genomics Institute. Published by Elsevier Ltd. All rights

reserved.

DOI: 10.1016/S1672-0229(10)60014-9

PMCID: PMC5054453

PMID: 20691398 [Indexed for MEDLINE]

1998. J Biomol NMR. 2010 Jun;47(2):85-99. doi: 10.1007/s10858-010-9407-y. Epub 2010 May

6.

A probabilistic approach for validating protein NMR chemical shift assignments.

Wang B(1), Wang Y, Wishart DS.

Author information:

(1)Shanghai American School Pudong, 201201, San Jia Gang, Pudong, Shanghai,

People's Republic of China.

It has been estimated that more than 20% of the proteins in the BMRB are

improperly referenced and that about 1% of all chemical shift assignments are

mis-assigned. These statistics also reflect the likelihood that any newly

assigned protein will have shift assignment or shift referencing errors. The

relatively high frequency of these errors continues to be a concern for the

biomolecular NMR community. While several programs do exist to detect and/or

correct chemical shift mis-referencing or chemical shift mis-assignments, most

can only do one, or the other. The one program (SHIFTCOR) that is capable of

handling both chemical shift mis-referencing and mis-assignments, requires the 3D

structure coordinates of the target protein. Given that chemical shift

mis-assignments and chemical shift re-referencing issues should ideally be

addressed prior to 3D structure determination, there is a clear need to develop a

structure-independent approach. Here, we present a new structure-independent

protocol, which is based on using residue-specific and secondary

structure-specific chemical shift distributions calculated over small (3-6

residue) fragments to identify mis-assigned resonances. The method is also able

to identify and re-reference mis-referenced chemical shift assignments.

Comparisons against existing re-referencing or mis-assignment detection programs

show that the method is as good or superior to existing approaches. The protocol

described here has been implemented into a freely available Java program called

"Probabilistic Approach for protein Nmr Assignment Validation (PANAV)" and as a

web server ( http://redpoll.pharmacy.ualberta.ca/PANAV ) which can be used to

validate and/or correct as well as re-reference assigned protein chemical shifts.

DOI: 10.1007/s10858-010-9407-y

PMID: 20446018 [Indexed for MEDLINE]

1999. Zebrafish. 2010 Jun;7(2):179-80. doi: 10.1089/zeb.2009.0624.

FishMap Zv8 update--a genomic regulatory map of zebrafish.

Bhartiya D(1), Maini J, Sharma M, Joshi P, Laddha SV, Jalali S, Patowary A,

Purkanti R, Lalwani M, Singh AR, Chauhan R, Singh N, Bhardwaj A, Scaria V,

Sivasubbu S.

Author information:

(1)Institute of Genomics and Integrative Biology (Council of Scientific and

Industrial Research), Delhi, India.

The advancements in genomics technologies and the amenability to large-scale

computational analysis have contributed immensely to the understanding of the

zebrafish genome, its organization, and its functional correlates. Translating

genomics information into biological meaning would require integration and

amenability of data and tools. FishMap is a community resource for genomic

datasets on zebrafish created with a vision to provide relevant and readily

available information to zebrafish researchers. The present update of FishMap has

kept up with the availability of the latest zebrafish genome assembly (Zv8). In

this update, particular emphasis has been given to noncoding RNAs and noncoding

RNA-mediated regulation in addition to genomic regulatory motifs, which are

emerging areas of vertebrate biology. FishMap Zv8 update also features a sequence

mapping and analysis server. Consistent with its commitment to make the

information freely available to the community, FishMap features options to share

data between compatible resources in addition to making it amenable to

programmatic access. FishMap Zv8 update is available at

http://fishmap2.igib.res.in.

DOI: 10.1089/zeb.2009.0624

PMID: 20528264 [Indexed for MEDLINE]

2000. BMC Res Notes. 2010 May 26;3:145. doi: 10.1186/1756-0500-3-145.

PRED\_PPI: a server for predicting protein-protein interactions based on sequence

data with probability assignment.

Guo Y(1), Li M, Pu X, Li G, Guang X, Xiong W, Li J.

Author information:

(1)College of Chemistry, Sichuan University, Chengdu 610064, PR China.

liml@scu.edu.cn.

BACKGROUND: Protein-protein interactions (PPIs) are crucial for almost all

cellular processes, including metabolic cycles, DNA transcription and

replication, and signaling cascades. Given the importance of PPIs, several

methods have been developed to detect them. Since the experimental methods are

time-consuming and expensive, developing computational methods for effectively

identifying PPIs is of great practical significance.

FINDINGS: Most previous methods were developed for predicting PPIs in only one

species, and do not account for probability estimations. In this work, a

relatively comprehensive prediction system was developed, based on a support

vector machine (SVM), for predicting PPIs in five organisms, specifically humans,

yeast, Drosophila, Escherichia coli, and Caenorhabditis elegans. This PPI

predictor includes the probability of its prediction in the output, so it can be

used to assess the confidence of each SVM prediction by the probability

assignment. Using a probability of 0.5 as the threshold for assigning class

labels, the method had an average accuracy for detecting protein interactions of

90.67% for humans, 88.99% for yeast, 90.09% for Drosophila, 92.73% for E. coli,

and 97.51% for C. elegans. Moreover, among the correctly predicted pairs, more

than 80% were predicted with a high probability of >/=0.8, indicating that this

tool could predict novel PPIs with high confidence.

CONCLUSIONS: Based on this work, a web-based system, Pred\_PPI, was constructed

for predicting PPIs from the five organisms. Users can predict novel PPIs and

obtain a probability value about the prediction using this tool. Pred\_PPI is

freely available at

http://cic.scu.edu.cn/bioinformatics/predict\_ppi/default.html.

DOI: 10.1186/1756-0500-3-145

PMCID: PMC2883990

PMID: 20500905

2001. J Theor Biol. 2010 May 21;264(2):326-33. doi: 10.1016/j.jtbi.2010.01.018. Epub

2010 Jan 20.

Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular

localization of Gram-negative bacterial proteins.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong

University, 800 Dongchuan Road, Shanghai 200240, China. hbshen@sjtu.edu.cn

By incorporating the information of gene ontology, functional domain, and

sequential evolution, a new predictor called Gneg-mPLoc was developed. It can be

used to identify Gram-negative bacterial proteins among the following eight

locations: (1) cytoplasm, (2) extracellular, (3) fimbrium, (4) flagellum, (5)

inner membrane, (6) nucleoid, (7) outer membrane, and (8) periplasm. It can also

be used to deal with the case when a query protein may simultaneously exist in

more than one location. Compared with the original predictor called Gneg-PLoc,

the new predictor is much more powerful and flexible. For a newly constructed

stringent benchmark dataset in which none of proteins included has >or=25%

pairwise sequence identity to any other in a same subset (location), the overall

jackknife success rate achieved by Gneg-mPLoc was 85.5%, which was more than 14%

higher than the corresponding rate by the Gneg-PLoc. As a user friendly

web-server, Gneg-mPLoc is freely accessible at

http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/.

Copyright (c) 2010 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2010.01.018

PMID: 20093124 [Indexed for MEDLINE]

2002. BMC Bioinformatics. 2010 May 20;11:268. doi: 10.1186/1471-2105-11-268.

Characterization of phylogenetic networks with NetTest.

Arenas M(1), Patricio M, Posada D, Valiente G.

Author information:

(1)Department of Biochemistry, Genetics and Immunology, University of Vigo,

E-36310 Vigo, Spain. miguelab@uvigo.es

BACKGROUND: Typical evolutionary events like recombination, hybridization or gene

transfer make necessary the use of phylogenetic networks to properly depict the

evolution of DNA and protein sequences. Although several theoretical classes have

been proposed to characterize these networks, they make stringent assumptions

that will likely not be met by the evolutionary process. We have recently shown

that the complexity of simulated networks is a function of the population

recombination rate, and that at moderate and large recombination rates the

resulting networks cannot be categorized. However, we do not know whether these

results extend to networks estimated from real data.

RESULTS: We introduce a web server for the categorization of explicit

phylogenetic networks, including the most relevant theoretical classes developed

so far. Using this tool, we analyzed statistical parsimony phylogenetic networks

estimated from approximately 5,000 DNA alignments, obtained from the NCBI PopSet

and Polymorphix databases. The level of characterization was correlated to

nucleotide diversity, and a high proportion of the networks derived from these

data sets could be formally characterized.

CONCLUSIONS: We have developed a public web server, NetTest (freely available

from the software section at http://darwin.uvigo.es), to formally characterize

the complexity of phylogenetic networks. Using NetTest we found that most

statistical parsimony networks estimated with the program TCS could be assigned

to a known network class. The level of network characterization was correlated to

nucleotide diversity and dependent upon the intra/interspecific levels, although

no significant differences were detected among genes. More research on the

properties of phylogenetic networks is clearly needed.

DOI: 10.1186/1471-2105-11-268

PMCID: PMC2880032

PMID: 20487540 [Indexed for MEDLINE]

2003. Bioinformatics. 2010 May 15;26(10):1299-307. doi: 10.1093/bioinformatics/btq114.

Epub 2010 Apr 12.

Quality assessment of protein model-structures using evolutionary conservation.

Kalman M(1), Ben-Tal N.

Author information:

(1)Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv

University, Ramat Aviv 69978, Israel.

MOTIVATION: Programs that evaluate the quality of a protein structural model are

important both for validating the structure determination procedure and for

guiding the model-building process. Such programs are based on properties of

native structures that are generally not expected for faulty models. One such

property, which is rarely used for automatic structure quality assessment, is the

tendency for conserved residues to be located at the structural core and for

variable residues to be located at the surface.

RESULTS: We present ConQuass, a novel quality assessment program based on the

consistency between the model structure and the protein's conservation pattern.

We show that it can identify problematic structural models, and that the scores

it assigns to the server models in CASP8 correlate with the similarity of the

models to the native structure. We also show that when the conservation

information is reliable, the method's performance is comparable and complementary

to that of the other single-structure quality assessment methods that

participated in CASP8 and that do not use additional structural information from

homologs.

AVAILABILITY: A perl implementation of the method, as well as the various perl

and R scripts used for the analysis are available at

http://bental.tau.ac.il/ConQuass/.

CONTACT: nirb@tauex.tau.ac.il

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq114

PMCID: PMC2865859

PMID: 20385730 [Indexed for MEDLINE]

2004. Bioinformatics. 2010 May 15;26(10):1370-1. doi: 10.1093/bioinformatics/btq137.

Epub 2010 Mar 30.

Studying the co-evolution of protein families with the Mirrortree web server.

Ochoa D(1), Pazos F.

Author information:

(1)National Centre for Biotechnology, Computational Systems Biology Group

(CNB-CSIC), c/ Darwin, 3. Cantoblanco, 28049 Madrid, Spain.

SUMMARY: The Mirrortree server allows to graphically and interactively study the

co-evolution of two protein families, and investigate their possible interactions

and functional relationships in a taxonomic context. The server includes the

possibility of starting from single sequences and hence it can be used by

non-expert users.

AVAILABILITY AND IMPLEMENTATION: The web server is freely available at

http://csbg.cnb.csic.es/mtserver. It was tested in the main web browsers. Adobe

Flash Player is required at the client side to perform the interactive assessment

of co-evolution.

CONTACT: pazos@cnb.csic.es

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btq137

PMID: 20363731 [Indexed for MEDLINE]

2005. Proteins. 2010 May 15;78(7):1789-97. doi: 10.1002/prot.22694.

Classification of transporters using efficient radial basis function networks

with position-specific scoring matrices and biochemical properties.

Ou YY(1), Chen SA, Gromiha MM.

Author information:

(1)Department of Computer Science and Engineering, Yuan Ze University, Chung-Li,

Taiwan.

Transporters are proteins that are involved in the movement of ions or molecules

across biological membranes. Transporters are generally classified into

channels/pores, electrochemical transporters, and active transporters.

Discriminating the specific class of transporters and their subfamilies are

essential tasks in computational biology for the advancement of structural and

functional genomics. We have systematically analyzed the amino acid composition,

residue pair preference and amino acid properties in six different families of

transporters. Utilizing the information, we have developed a radial basis

function (RBF) network method based on profiles obtained with position specific

scoring matrices for discriminating transporters belonging to three different

classes and six families. Our method showed a fivefold cross validation accuracy

of 76%, 73%, and 69% for discriminating transporters and nontransporters, three

different classes and six different families of transporters, respectively.

Further, the method was tested with independent datasets, which showed similar

level of accuracy. A web server has been developed for discriminating

transporters based on three classes and six families, and it is available at

http://rbf.bioinfo.tw/ approximately sachen/tcrbf.html. We suggest that our

method could be effectively used to identify transporters and discriminating them

into different classes and families.

Proteins 2010;. (c) 2010 Wiley-Liss, Inc.

DOI: 10.1002/prot.22694

PMID: 20196081 [Indexed for MEDLINE]

2006. Proteins. 2010 May 15;78(7):1618-30. doi: 10.1002/prot.22678.

Circular permuted proteins in the universe of protein folds.

Schmidt-Goenner T(1), Guerler A, Kolbeck B, Knapp EW.

Author information:

(1)Institute of Chemistry and Biochemistry, Freie Universität Berlin,

Fabeckstrasse 36a, 14195 Berlin, Germany.

Finding and identifying circular permuted protein pairs (CPP) is one of the

harder tasks for structure alignment programs, because of the different location

of the break in the polypeptide chain connectivity. The protein structure

alignment tool GANGSTA+ was used to search for CPPs in a database of nearly

10,000 protein structures. It also allows determination of the statistical

significance of the occurrence of circular permutations in the protein universe.

The number of detected CPPs was found to be higher than expected, raising

questions about the evolutionary processes leading to CPPs. The GANGSTA+ protein

structure alignment tool is available online via the web server at

http://gangsta.chemie.fu-berlin.de. On the same webpage the complete data base of

similar protein structure pairs based on the ASTRAL40 set of protein domains is

provided and one can select CPPs specifically.

Proteins 2010. (c) 2009 Wiley-Liss, Inc.

DOI: 10.1002/prot.22678

PMID: 20112421 [Indexed for MEDLINE]

2007. BMC Med Inform Decis Mak. 2010 May 13;10:26. doi: 10.1186/1472-6947-10-26.

From design to implementation--the Joint Asia Diabetes Evaluation (JADE) program:

a descriptive report of an electronic web-based diabetes management program.

Ko GT(1), So WY, Tong PC, Le Coguiec F, Kerr D, Lyubomirsky G, Tamesis B,

Wolthers T, Nan J, Chan J.

Author information:

(1)Asia Diabetes Foundation, Flat 4B, Block B, Prince of Wales Hospital, Shatin,

Hong Kong SAR, China.

BACKGROUND: The Joint Asia Diabetes Evaluation (JADE) Program is a web-based

program incorporating a comprehensive risk engine, care protocols, and clinical

decision support to improve ambulatory diabetes care.

METHODS: The JADE Program uses information technology to facilitate healthcare

professionals to create a diabetes registry and to deliver an evidence-based care

and education protocol tailored to patients' risk profiles. With written informed

consent from participating patients and care providers, all data are anonymized

and stored in a databank to establish an Asian Diabetes Database for research and

publication purpose.

RESULTS: The JADE electronic portal (e-portal: http://www.jade-adf.org) is

implemented as a Java application using the Apache web server, the mySQL database

and the Cocoon framework. The JADE e-portal comprises a risk engine which

predicts 5-year probability of major clinical events based on parameters

collected during an annual comprehensive assessment. Based on this risk

stratification, the JADE e-portal recommends a care protocol tailored to these

risk levels with decision support triggered by various risk factors. Apart from

establishing a registry for quality assurance and data tracking, the JADE

e-portal also displays trends of risk factor control at each visit to promote

doctor-patient dialogues and to empower both parties to make informed decisions.

CONCLUSIONS: The JADE Program is a prototype using information technology to

facilitate implementation of a comprehensive care model, as recommended by the

International Diabetes Federation. It also enables health care teams to record,

manage, track and analyze the clinical course and outcomes of people with

diabetes.

DOI: 10.1186/1472-6947-10-26

PMCID: PMC2876072

PMID: 20465815 [Indexed for MEDLINE]

2008. BMC Genomics. 2010 May 11;11:293. doi: 10.1186/1471-2164-11-293.

Ensembl variation resources.

Chen Y(1), Cunningham F, Rios D, McLaren WM, Smith J, Pritchard B, Spudich GM,

Brent S, Kulesha E, Marin-Garcia P, Smedley D, Birney E, Flicek P.

Author information:

(1)European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton,

Cambridge CB10 1SD, UK.

BACKGROUND: The maturing field of genomics is rapidly increasing the number of

sequenced genomes and producing more information from those previously sequenced.

Much of this additional information is variation data derived from sampling

multiple individuals of a given species with the goal of discovering new variants

and characterising the population frequencies of the variants that are already

known. These data have immense value for many studies, including those designed

to understand evolution and connect genotype to phenotype. Maximising the utility

of the data requires that it be stored in an accessible manner that facilitates

the integration of variation data with other genome resources such as gene

annotation and comparative genomics.

DESCRIPTION: The Ensembl project provides comprehensive and integrated variation

resources for a wide variety of chordate genomes. This paper provides a detailed

description of the sources of data and the methods for creating the Ensembl

variation databases. It also explores the utility of the information by

explaining the range of query options available, from using interactive web

displays, to online data mining tools and connecting directly to the data servers

programmatically. It gives a good overview of the variation resources and future

plans for expanding the variation data within Ensembl.

CONCLUSIONS: Variation data is an important key to understanding the functional

and phenotypic differences between individuals. The development of new sequencing

and genotyping technologies is greatly increasing the amount of variation data

known for almost all genomes. The Ensembl variation resources are integrated into

the Ensembl genome browser and provide a comprehensive way to access this data in

the context of a widely used genome bioinformatics system. All Ensembl data is

freely available at http://www.ensembl.org and from the public MySQL database

server at ensembldb.ensembl.org.

DOI: 10.1186/1471-2164-11-293

PMCID: PMC2894800

PMID: 20459805 [Indexed for MEDLINE]

2009. BMC Genomics. 2010 May 10;11:290. doi: 10.1186/1471-2164-11-290.

Analysis of gene evolution and metabolic pathways using the Candida Gene Order

Browser.

Fitzpatrick DA(1), O'Gaora P, Byrne KP, Butler G.

Author information:

(1)UCD School of Biomolecular and Biomedical Science, Conway Institute,

University College Dublin, Belfield, Dublin 4, Ireland. david.fitzpatrick@nuim.ie

BACKGROUND: Candida species are the most common cause of opportunistic fungal

infection worldwide. Recent sequencing efforts have provided a wealth of Candida

genomic data. We have developed the Candida Gene Order Browser (CGOB), an online

tool that aids comparative syntenic analyses of Candida species. CGOB

incorporates all available Candida clade genome sequences including two Candida

albicans isolates (SC5314 and WO-1) and 8 closely related species (Candida

dubliniensis, Candida tropicalis, Candida parapsilosis, Lodderomyces

elongisporus, Debaryomyces hansenii, Pichia stipitis, Candida guilliermondii and

Candida lusitaniae). Saccharomyces cerevisiae is also included as a reference

genome.

RESULTS: CGOB assignments of homology were manually curated based on sequence

similarity and synteny. In total CGOB includes 65617 genes arranged into 13625

homology columns. We have also generated improved Candida gene sets by

merging/removing partial genes in each genome. Interrogation of CGOB revealed

that the majority of tandemly duplicated genes are under strong purifying

selection in all Candida species. We identified clusters of adjacent genes

involved in the same metabolic pathways (such as catabolism of biotin, galactose

and N-acetyl glucosamine) and we showed that some clusters are species or

lineage-specific. We also identified one example of intron gain in C. albicans.

CONCLUSIONS: Our analysis provides an important resource that is now available

for the Candida community. CGOB is available at http://cgob.ucd.ie.

DOI: 10.1186/1471-2164-11-290

PMCID: PMC2880306

PMID: 20459735 [Indexed for MEDLINE]

2010. J Theor Biol. 2010 May 7;264(1):130-5. doi: 10.1016/j.jtbi.2010.01.013. Epub 2010

Jan 18.

Lysine acetylation sites prediction using an ensemble of support vector machine

classifiers.

Xu Y(1), Wang XB, Ding J, Wu LY, Deng NY.

Author information:

(1)College of Science, China Agricultural University, Beijing 100083, China.

Lysine acetylation is an essentially reversible and high regulated

post-translational modification which regulates diverse protein properties.

Experimental identification of acetylation sites is laborious and expensive.

Hence, there is significant interest in the development of computational methods

for reliable prediction of acetylation sites from amino acid sequences. In this

paper we use an ensemble of support vector machine classifiers to perform this

work. The experimentally determined acetylation lysine sites are extracted from

Swiss-Prot database and scientific literatures. Experiment results show that an

ensemble of support vector machine classifiers outperforms single support vector

machine classifier and other computational methods such as PAIL and LysAcet on

the problem of predicting acetylation lysine sites. The resulting method has been

implemented in EnsemblePail, a web server for lysine acetylation sites prediction

available at http://www.aporc.org/EnsemblePail/.

Copyright (c) 2010 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.jtbi.2010.01.013

PMID: 20085770 [Indexed for MEDLINE]

2011. BMC Bioinformatics. 2010 May 6;11:231. doi: 10.1186/1471-2105-11-231.

RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search

the three-dimensional fragments within RNA structures.

Popenda M(1), Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, Adamiak

RW.

Author information:

(1)Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland.

BACKGROUND: Recent discoveries concerning novel functions of RNA, such as RNA

interference, have contributed towards the growing importance of the field. In

this respect, a deeper knowledge of complex three-dimensional RNA structures is

essential to understand their new biological functions. A number of bioinformatic

tools have been proposed to explore two major structural databases (PDB, NDB) in

order to analyze various aspects of RNA tertiary structures. One of these tools

is RNA FRABASE 1.0, the first web-accessible database with an engine for

automatic search of 3D fragments within PDB-derived RNA structures. This search

is based upon the user-defined RNA secondary structure pattern. In this paper, we

present and discuss RNA FRABASE 2.0. This second version of the system represents

a major extension of this tool in terms of providing new data and a wide spectrum

of novel functionalities. An intuitionally operated web server platform enables

very fast user-tailored search of three-dimensional RNA fragments, their

multi-parameter conformational analysis and visualization.

DESCRIPTION: RNA FRABASE 2.0 has stored information on 1565 PDB-deposited RNA

structures, including all NMR models. The RNA FRABASE 2.0 search engine

algorithms operate on the database of the RNA sequences and the new library of

RNA secondary structures, coded in the dot-bracket format extended to hold

multi-stranded structures and to cover residues whose coordinates are missing in

the PDB files. The library of RNA secondary structures (and their graphics) is

made available. A high level of efficiency of the 3D search has been achieved by

introducing novel tools to formulate advanced searching patterns and to screen

highly populated tertiary structure elements. RNA FRABASE 2.0 also stores data

and conformational parameters in order to provide "on the spot" structural

filters to explore the three-dimensional RNA structures. An instant visualization

of the 3D RNA structures is provided. RNA FRABASE 2.0 is freely available at

http://rnafrabase.cs.put.poznan.pl.

CONCLUSIONS: RNA FRABASE 2.0 provides a novel database and powerful search engine

which is equipped with new data and functionalities that are unavailable

elsewhere. Our intention is that this advanced version of the RNA FRABASE will be

of interest to all researchers working in the RNA field.

DOI: 10.1186/1471-2105-11-231

PMCID: PMC2873543

PMID: 20459631 [Indexed for MEDLINE]

2012. BMC Bioinformatics. 2010 May 6;11:230. doi: 10.1186/1471-2105-11-230.

BISMA--fast and accurate bisulfite sequencing data analysis of individual clones

from unique and repetitive sequences.

Rohde C(1), Zhang Y, Reinhardt R, Jeltsch A.

Author information:

(1)School of Engineering and Science, Jacobs University Bremen, Campus Ring 1,

28725 Bremen, Germany.

BACKGROUND: Bisulfite sequencing is a popular method to analyze DNA methylation

patterns at high resolution. A region of interest is targeted by PCR and about

20-50 subcloned DNA molecules are usually analyzed, to determine the methylation

status at single CpG sites and molecule resolution.

RESULTS: The BISMA (Bisulfite Sequencing DNA Methylation Analysis) software for

analysis of primary bisulfite sequencing data implements sequencing data

extraction and enhanced data processing, quality controls, analysis and

presentation of the methylation state. It uses an improved strategy for detection

of clonal molecules and accurate CpG site detection and it supports for the first

time analysis of repetitive sequences.

CONCLUSIONS: BISMA works highly automated but still provides the user full

control over all steps of the analysis. The BISMA software is freely available as

an online tool for academic purposes for the analysis of bisulfite sequencing

data from both unique and repetitive sequences

http://biochem.jacobs-university.de/BDPC/BISMA/.

DOI: 10.1186/1471-2105-11-230

PMCID: PMC2877691

PMID: 20459626 [Indexed for MEDLINE]

2013. Bioinformatics. 2010 May 1;26(9):1258-9. doi: 10.1093/bioinformatics/btq116. Epub

2010 Mar 17.

SeSAW: balancing sequence and structural information in protein functional

mapping.

Standley DM(1), Yamashita R, Kinjo AR, Toh H, Nakamura H.

Author information:

(1)WPI Immunology Frontier Research Center (IFReC), Osaka University, Suita,

Osaka, Japan. standley@ifrec.osaka-u.ac.jp

MOTIVATION: Functional similarity between proteins is evident at both the

sequence and structure levels. SeSAW is a web-based program for identifying

functionally or evolutionarily conserved motifs in protein structures by locating

sequence and structural similarities, and quantifying these at the level of

individual residues. Results can be visualized in 2D, as annotated alignments, or

in 3D, as structural superpositions. An example is given for both an

experimentally determined query structure and a homology model.

AVAILABILITY AND IMPLEMENTATION: The web server is located at

http://www.pdbj.org/SeSAW/.

DOI: 10.1093/bioinformatics/btq116

PMCID: PMC2859130

PMID: 20299324 [Indexed for MEDLINE]

2014. Biotechniques. 2010 May;48(5):405-8. doi: 10.2144/000113370.

DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and

informative sites.

Deng W(1), Maust BS, Nickle DC, Learn GH, Liu Y, Heath L, Kosakovsky Pond SL,

Mullins JI.

Author information:

(1)Department of Microbiology, University of Washington School of Medicine,

Seattle, WA 98195, USA.

DIVEIN is a web interface that performs automated phylogenetic and other analyses

of nucleotide and amino acid sequences. Starting with a set of aligned sequences,

DIVEIN estimates evolutionary parameters and phylogenetic trees while allowing

the user to choose from a variety of evolutionary models; it then reconstructs

the consensus (CON), most recent common ancestor (MRCA), and center of tree (COT)

sequences. DIVEIN also provides tools for further analyses, including condensing

sequence alignments to show only informative sites or private mutations;

computing phylogenetic or pairwise divergence from any user-specified sequence

(CON, MRCA, COT, or existing sequence from the alignment); computing and

outputting all genetic distances in column format; calculating summary statistics

of diversity and divergence from pairwise distances; and graphically representing

the inferred tree and plots of divergence, diversity, and distance distribution

histograms. DIVEIN is available at

http://indra.mullins.microbiol.washington.edu/DIVEIN.

DOI: 10.2144/000113370

PMCID: PMC3133969

PMID: 20569214 [Indexed for MEDLINE]

2015. Nucleic Acids Res. 2010 May;38(8):2617-23. doi: 10.1093/nar/gkq093. Epub 2010 Mar

7.

CLONEQC: lightweight sequence verification for synthetic biology.

Lee PA(1), Dymond JS, Scheifele LZ, Richardson SM, Foelber KJ, Boeke JD, Bader

JS.

Author information:

(1)Department of Computer Science, Johns Hopkins University, 3400 N. Charles St.,

Baltimore, MD 21215, USA.

Synthetic biology projects aim to produce physical DNA that matches a designed

target sequence. Chemically synthesized oligomers are generally used as the

starting point for building larger and larger sequences. Due to the error rate of

chemical synthesis, these oligomers can have many differences from the target

sequence. As oligomers are joined together to make larger and larger synthetic

intermediates, it becomes essential to perform quality control to eliminate

intermediates with errors and retain only those DNA molecules that are error free

with respect to the target. This step is often performed by transforming bacteria

with synthetic DNA and sequencing colonies until a clone with a perfect sequence

is identified. Here we present CloneQC, a lightweight software pipeline available

as a free web server and as source code that performs quality control on

sequenced clones. Input to the server is a list of desired sequences and forward

and reverse reads for each clone. The server generates summary statistics (error

rates and success rates target-by-target) and a detailed report of perfect

clones. This software will be useful to laboratories conducting in-house DNA

synthesis and is available at http://cloneqc.thruhere.net/ and as Berkeley

Software Distribution (BSD) licensed source.

DOI: 10.1093/nar/gkq093

PMCID: PMC2860120

PMID: 20211841 [Indexed for MEDLINE]

2016. Proteins. 2010 May 1;78(6):1503-19. doi: 10.1002/prot.22668.

FiberDock: Flexible induced-fit backbone refinement in molecular docking.

Mashiach E(1), Nussinov R, Wolfson HJ.

Author information:

(1)Tel Aviv University, Israel.

Upon binding, proteins undergo conformational changes. These changes often

prevent rigid-body docking methods from predicting the 3D structure of a complex

from the unbound conformations of its proteins. Handling protein backbone

flexibility is a major challenge for docking methodologies, as backbone

flexibility adds a huge number of degrees of freedom to the search space, and

therefore considerably increases the running time of docking algorithms. Normal

mode analysis permits description of protein flexibility as a linear combination

of discrete movements (modes). Low-frequency modes usually describe the

large-scale conformational changes of the protein. Therefore, many docking

methods model backbone flexibility by using only few modes, which have the lowest

frequencies. However, studies show that due to molecular interactions, many

proteins also undergo local and small-scale conformational changes, which are

described by high-frequency normal modes. Here we present a new method,

FiberDock, for docking refinement which models backbone flexibility by an

unlimited number of normal modes. The method iteratively minimizes the structure

of the flexible protein along the most relevant modes. The relevance of a mode is

calculated according to the correlation between the chemical forces, applied on

each atom, and the translation vector of each atom, according to the normal mode.

The results show that the method successfully models backbone movements that

occur during molecular interactions and considerably improves the accuracy and

the ranking of rigid-docking models of protein-protein complexes. A web server

for the FiberDock method is available at:

http://bioinfo3d.cs.tau.ac.il/FiberDock.

2009 Wiley-Liss, Inc.

DOI: 10.1002/prot.22668

PMCID: PMC4290165

PMID: 20077569 [Indexed for MEDLINE]

2017. Syst Biol. 2010 May;59(3):307-21. doi: 10.1093/sysbio/syq010. Epub 2010 Mar 29.

New algorithms and methods to estimate maximum-likelihood phylogenies: assessing

the performance of PhyML 3.0.

Guindon S(1), Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.

Author information:

(1)Méthodes et Algorithmes pour la Bioinformatique, LIRMM, Centre National de la

Recherche Scientifique, Université de Montpellier, Montpellier Cedex 5, France.

PhyML is a phylogeny software based on the maximum-likelihood principle. Early

PhyML versions used a fast algorithm performing nearest neighbor interchanges to

improve a reasonable starting tree topology. Since the original publication

(Guindon S., Gascuel O. 2003. A simple, fast and accurate algorithm to estimate

large phylogenies by maximum likelihood. Syst. Biol. 52:696-704), PhyML has been

widely used (>2500 citations in ISI Web of Science) because of its simplicity and

a fair compromise between accuracy and speed. In the meantime, research around

PhyML has continued, and this article describes the new algorithms and methods

implemented in the program. First, we introduce a new algorithm to search the

tree space with user-defined intensity using subtree pruning and regrafting

topological moves. The parsimony criterion is used here to filter out the least

promising topology modifications with respect to the likelihood function. The

analysis of a large collection of real nucleotide and amino acid data sets of

various sizes demonstrates the good performance of this method. Second, we

describe a new test to assess the support of the data for internal branches of a

phylogeny. This approach extends the recently proposed approximate

likelihood-ratio test and relies on a nonparametric, Shimodaira-Hasegawa-like

procedure. A detailed analysis of real alignments sheds light on the links

between this new approach and the more classical nonparametric bootstrap method.

Overall, our tests show that the last version (3.0) of PhyML is fast, accurate,

stable, and ready to use. A Web server and binary files are available from

http://www.atgc-montpellier.fr/phyml/.

DOI: 10.1093/sysbio/syq010

PMID: 20525638 [Indexed for MEDLINE]

2018. BMC Bioinformatics. 2010 Apr 29;11 Suppl 3:S4. doi: 10.1186/1471-2105-11-S3-S4.

A conformal Bayesian network for classification of Mycobacterium tuberculosis

complex lineages.

Aminian M(1), Shabbeer A, Bennett KP.

Author information:

(1)Departments of Mathematical Science and Computer Science, Rensselaer

Polytechnic Institute, Troy, New York, USA. aminim@cs.rpi.edu

BACKGROUND: We present a novel conformal Bayesian network (CBN) to classify

strains of Mycobacterium tuberculosis Complex (MTBC) into six major genetic

lineages based on two high-throuput biomarkers: mycobacterial interspersed

repetitive units (MIRU) and spacer oligonucleotide typing (spoligotyping). MTBC

is the causative agent of tuberculosis (TB), which remains one of the leading

causes of disease and morbidity world-wide. DNA fingerprinting methods such as

MIRU and spoligotyping are key components in the control and tracking of modern

TB.

RESULTS: CBN is designed to exploit background knowledge about MTBC biomarkers.

It can be trained on large historical TB databases of various subsets of MTBC

biomarkers. During TB control efforts not all biomarkers may be available. So,

CBN is designed to predict the major lineage of isolates genotyped by any

combination of the PCR-based typing methods: spoligotyping and MIRU typing. CBN

achieves high accuracy on three large MTBC collections consisting of over 34,737

isolates genotyped by different combinations of spoligotypes, 12 loci of MIRU,

and 24 loci of MIRU. CBN captures distinct MIRU and spoligotype signatures

associated with each lineage, explaining its excellent performance. Visualization

of MIRU and spoligotype signatures yields insight into both how the model works

and the genetic diversity of MTBC.

CONCLUSIONS: CBN conforms to the available PCR-based biological markers and

achieves high performance in identifying major lineages of MTBC. The method can

be readily extended as new biomarkers are introduced for TB tracking and control.

An online tool (http://www.cs.rpi.edu/~bennek/tbinsight/tblineage) makes the CBN

model available for TB control and research efforts.

DOI: 10.1186/1471-2105-11-S3-S4

PMCID: PMC2863063

PMID: 20438651 [Indexed for MEDLINE]

2019. BMC Bioinformatics. 2010 Apr 29;11 Suppl 3:S1. doi: 10.1186/1471-2105-11-S3-S1.

SeqRate: sequence-based protein folding type classification and rates prediction.

Lin GN(1), Wang Z, Xu D, Cheng J.

Author information:

(1)Informatics Institute, University of Missouri, Columbia, 65211, USA.

guanlin@mail.missouri.edu

BACKGROUND: Protein folding rate is an important property of a protein.

Predicting protein folding rate is useful for understanding protein folding

process and guiding protein design. Most previous methods of predicting protein

folding rate require the tertiary structure of a protein as an input. And most

methods do not distinguish the different kinetic nature (two-state folding or

multi-state folding) of the proteins. Here we developed a method, SeqRate, to

predict both protein folding kinetic type (two-state versus multi-state) and

real-value folding rate using sequence length, amino acid composition, contact

order, contact number, and secondary structure information predicted from only

protein sequence with support vector machines.

RESULTS: We systematically studied the contributions of individual features to

folding rate prediction. On a standard benchmark dataset, the accuracy of folding

kinetic type classification is 80%. The Pearson correlation coefficient and the

mean absolute difference between predicted and experimental folding rates (sec-1)

in the base-10 logarithmic scale are 0.81 and 0.79 for two-state protein folders,

and 0.80 and 0.68 for three-state protein folders. SeqRate is the first

sequence-based method for protein folding type classification and its accuracy of

fold rate prediction is improved over previous sequence-based methods. Its

performance can be further enhanced with additional information, such as

structure-based geometric contacts, as inputs.

CONCLUSIONS: Both the web server and software of predicting folding rate are

publicly available at http://casp.rnet.missouri.edu/fold\_rate/index.html.

DOI: 10.1186/1471-2105-11-S3-S1

PMCID: PMC2863059

PMID: 20438647 [Indexed for MEDLINE]

2020. Bioinformatics. 2010 Apr 15;26(8):1125-6. doi: 10.1093/bioinformatics/btq086.

Epub 2010 Mar 1.

ScripTree: scripting phylogenetic graphics.

Chevenet F(1), Croce O, Hebrard M, Christen R, Berry V.

Author information:

(1)GEMI, IRD/CNRS UMR 2724, 911 avenue Agropolis, 34394 Montpellier, France.

chevenet@ird.fr

There is a large amount of tools for interactive display of phylogenetic trees.

However, there is a shortage of tools for the automation of tree rendering.

Scripting phylogenetic graphics would enable the saving of graphical analyses

involving numerous and complex tree handling operations and would allow the

automation of repetitive tasks. ScripTree is a tool intended to fill this gap. It

is an interpreter to be used in batch mode. Phylogenetic graphics instructions,

related to tree rendering as well as tree annotation, are stored in a text file

and processed in a sequential way.AVAILABILITY: ScripTree can be used online or

downloaded at www.scriptree.org, under the GPL license.

IMPLEMENTATION: ScripTree, written in Tcl/Tk, is a cross-platform application

available for Windows and Unix-like systems including OS X. It can be used either

as a stand-alone package or included in a bioinformatic pipeline and linked to a

HTTP server.

DOI: 10.1093/bioinformatics/btq086

PMID: 20194627 [Indexed for MEDLINE]

2021. Bioinformatics. 2010 Apr 15;26(8):1122-4. doi: 10.1093/bioinformatics/btq090.

Epub 2010 Mar 1.

An Ergatis-based prokaryotic genome annotation web server.

Hemmerich C(1), Buechlein A, Podicheti R, Revanna KV, Dong Q.

Author information:

(1)Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN

47405, USA.

SUMMARY: Ergatis is a flexible workflow management system for designing and

executing complex bioinformatics pipelines. However, its complexity restricts its

usage to only highly skilled bioinformaticians. We have developed a web-based

prokaryotic genome annotation server, Integrative Services for Genomics Analysis

(ISGA), which builds upon the Ergatis workflow system, integrates other dynamic

analysis tools and provides intuitive web interfaces for biologists to customize

and execute their own annotation pipelines. ISGA is designed to be installed at

genomics core facilities and be used directly by biologists.

AVAILABILITY: ISGA is accessible at http://isga.cgb.indiana.edu/ and the system

is also freely available for local installation.

DOI: 10.1093/bioinformatics/btq090

PMID: 20194626 [Indexed for MEDLINE]

2022. BMC Bioinformatics. 2010 Apr 13;11:187. doi: 10.1186/1471-2105-11-187.

Artificial and natural duplicates in pyrosequencing reads of metagenomic data.

Niu B(1), Fu L, Sun S, Li W.

Author information:

(1)California Institute for Telecommunications and Information Technology,

University of California San Diego, La Jolla, California 92093, USA.

BACKGROUND: Artificial duplicates from pyrosequencing reads may lead to incorrect

interpretation of the abundance of species and genes in metagenomic studies.

Duplicated reads were filtered out in many metagenomic projects. However, since

the duplicated reads observed in a pyrosequencing run also include natural

(non-artificial) duplicates, simply removing all duplicates may also cause

underestimation of abundance associated with natural duplicates.

RESULTS: We implemented a method for identification of exact and nearly identical

duplicates from pyrosequencing reads. This method performs an all-against-all

sequence comparison and clusters the duplicates into groups using an algorithm

modified from our previous sequence clustering method cd-hit. This method can

process a typical dataset in approximately 10 minutes; it also provides a

consensus sequence for each group of duplicates. We applied this method to the

underlying raw reads of 39 genomic projects and 10 metagenomic projects that

utilized pyrosequencing technique. We compared the occurrences of the duplicates

identified by our method and the natural duplicates made by independent

simulations. We observed that the duplicates, including both artificial and

natural duplicates, make up 4-44% of reads. The number of natural duplicates

highly correlates with the samples' read density (number of reads divided by

genome size). For high-complexity metagenomic samples lacking dominant species,

natural duplicates only make up <1% of all duplicates. But for some other samples

like transcriptomic samples, majority of the observed duplicates might be natural

duplicates.

CONCLUSIONS: Our method is available from http://cd-hit.org as a downloadable

program and a web server. It is important not only to identify the duplicates

from metagenomic datasets but also to distinguish whether they are artificial or

natural duplicates. We provide a tool to estimate the number of natural

duplicates according to user-defined sample types, so users can decide whether to

retain or remove duplicates in their projects.

DOI: 10.1186/1471-2105-11-187

PMCID: PMC2874554

PMID: 20388221 [Indexed for MEDLINE]

2023. BMC Bioinformatics. 2010 Apr 12;11:185. doi: 10.1186/1471-2105-11-185.

web cellHTS2: a web-application for the analysis of high-throughput screening

data.

Pelz O(1), Gilsdorf M, Boutros M.

Author information:

(1)German Cancer Research Center (DKFZ), Div Signaling and Functional Genomics

and University of Heidelberg, Faculty of Medicine Mannheim, Dept Cell and

Molecular Biology, Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany.

BACKGROUND: The analysis of high-throughput screening data sets is an expanding

field in bioinformatics. High-throughput screens by RNAi generate large primary

data sets which need to be analyzed and annotated to identify relevant phenotypic

hits. Large-scale RNAi screens are frequently used to identify novel factors that

influence a broad range of cellular processes, including signaling pathway

activity, cell proliferation, and host cell infection. Here, we present a

web-based application utility for the end-to-end analysis of large cell-based

screening experiments by cellHTS2.

RESULTS: The software guides the user through the configuration steps that are

required for the analysis of single or multi-channel experiments. The

web-application provides options for various standardization and normalization

methods, annotation of data sets and a comprehensive HTML report of the screening

data analysis, including a ranked hit list. Sessions can be saved and restored

for later re-analysis. The web frontend for the cellHTS2 R/Bioconductor package

interacts with it through an R-server implementation that enables highly parallel

analysis of screening data sets. web cellHTS2 further provides a file import and

configuration module for common file formats.

CONCLUSIONS: The implemented web-application facilitates the analysis of

high-throughput data sets and provides a user-friendly interface. web cellHTS2 is

accessible online at http://web-cellHTS2.dkfz.de. A standalone version as a

virtual appliance and source code for platforms supporting Java 1.5.0 can be

downloaded from the web cellHTS2 page. web cellHTS2 is freely distributed under

GPL.

DOI: 10.1186/1471-2105-11-185

PMCID: PMC3098057

PMID: 20385013 [Indexed for MEDLINE]

2024. BMC Evol Biol. 2010 Apr 12;10:99. doi: 10.1186/1471-2148-10-99.

FLU, an amino acid substitution model for influenza proteins.

Dang CC(1), Le QS, Gascuel O, Le VS.

Author information:

(1)College of Technology, Vietnam National University Hanoi, Cau Giay, Hanoi,

Vietnam.

BACKGROUND: The amino acid substitution model is the core component of many

protein analysis systems such as sequence similarity search, sequence alignment,

and phylogenetic inference. Although several general amino acid substitution

models have been estimated from large and diverse protein databases, they remain

inappropriate for analyzing specific species, e.g., viruses. Emerging epidemics

of influenza viruses raise the need for comprehensive studies of these dangerous

viruses. We propose an influenza-specific amino acid substitution model to

enhance the understanding of the evolution of influenza viruses.

RESULTS: A maximum likelihood approach was applied to estimate an amino acid

substitution model (FLU) from approximately 113,000 influenza protein sequences,

consisting of approximately 20 million residues. FLU outperforms 14 widely used

models in constructing maximum likelihood phylogenetic trees for the majority of

influenza protein alignments. On average, FLU gains approximately 42 log

likelihood points with an alignment of 300 sites. Moreover, topologies of trees

constructed using FLU and other models are frequently different. FLU does indeed

have an impact on likelihood improvement as well as tree topologies. It was

implemented in PhyML and can be downloaded from

ftp://ftp.sanger.ac.uk/pub/1000genomes/lsq/FLU or included in PhyML 3.0 server at

http://www.atgc-montpellier.fr/phyml/.

CONCLUSIONS: FLU should be useful for any influenza protein analysis system which

requires an accurate description of amino acid substitutions.

DOI: 10.1186/1471-2148-10-99

PMCID: PMC2873421

PMID: 20384985 [Indexed for MEDLINE]

2025. Bioinformatics. 2010 Apr 1;26(7):969-70. doi: 10.1093/bioinformatics/btq068. Epub

2010 Feb 19.

CandiSNPer: a web tool for the identification of candidate SNPs for causal

variants.

Schmitt AO(1), Assmus J, Bortfeldt RH, Brockmann GA.

Author information:

(1)Department for Crop and Animal Sciences, Humboldt-Universität zu Berlin,

Invalidenstrasse 42, 10115 Berlin, Germany. armin.schmitt@agrar.hu-berlin.de

SUMMARY: Human single nucleotide polymorphism (SNP) chips which are used in

genome-wide association studies (GWAS) permit the genotyping of up to 4 million

SNPs simultaneously. To date, about 1000 human SNPs have been identified as

statistically significantly associated with a disease or another trait of

interest. The identified SNP is not necessarily the causal variant, but it is

rather in linkage disequilibrium (LD) with it. CandiSNPer is a software tool that

determines the LD region around a significant SNP from a GWAS. It provides a list

with functional annotation and LD values for the SNPs found in the LD region.

This list contains not only the SNPs for which genotyping data are available, but

all SNPs with rs-IDs, thus increasing the likelihood to include the causal

variant. Furthermore, plots showing the LD values are generated. CandiSNPer

facilitates the preselection of candidate SNPs for causal variants.

AVAILABILITY AND IMPLEMENTATION: The CandiSNPer server is freely available at

http://www2.hu-berlin.de/wikizbnutztier/software/CandiSNPer. The source code is

available to academic users 'as is' upon request. The web site is implemented in

Perl and R and runs on an Apache server. The Ensembl database is queried for SNP

data via Perl APIs.

DOI: 10.1093/bioinformatics/btq068

PMID: 20172942 [Indexed for MEDLINE]

2026. Bioinformatics. 2010 Apr 1;26(7):882-8. doi: 10.1093/bioinformatics/btq058. Epub

2010 Feb 11.

MULTICOM: a multi-level combination approach to protein structure prediction and

its assessments in CASP8.

Wang Z(1), Eickholt J, Cheng J.

Author information:

(1)Department of Computer Science, Informatics Institute and C. Bond Life Science

Center, University of Missouri, Columbia, MO 65211, USA.

MOTIVATION: Protein structure prediction is one of the most important problems in

structural bioinformatics. Here we describe MULTICOM, a multi-level combination

approach to improve the various steps in protein structure prediction. In

contrast to those methods which look for the best templates, alignments and

models, our approach tries to combine complementary and alternative templates,

alignments and models to achieve on average better accuracy.

RESULTS: The multi-level combination approach was implemented via five automated

protein structure prediction servers and one human predictor which participated

in the eighth Critical Assessment of Techniques for Protein Structure Prediction

(CASP8), 2008. The MULTICOM servers and human predictor were consistently ranked

among the top predictors on the CASP8 benchmark. The methods can predict

moderate- to high-resolution models for most template-based targets and

low-resolution models for some template-free targets. The results show that the

multi-level combination of complementary templates, alternative alignments and

similar models aided by model quality assessment can systematically improve both

template-based and template-free protein modeling.

AVAILABILITY: The MULTICOM server is freely available at

http://casp.rnet.missouri.edu/multicom\_3d.html .

DOI: 10.1093/bioinformatics/btq058

PMCID: PMC2844995

PMID: 20150411 [Indexed for MEDLINE]

2027. Comput Biol Chem. 2010 Apr;34(2):126-30. doi: 10.1016/j.compbiolchem.2010.03.006.

Epub 2010 Apr 3.

ProSTRIP: A method to find similar structural repeats in three-dimensional

protein structures.

Sabarinathan R(1), Basu R, Sekar K.

Author information:

(1)Bioinformatics Centre (Centre of Excellence in Structural Biology and

Bio-computing), Indian Institute of Science, Bangalore 560 012, India.

The occurrence of similar structural repeats in a protein structure has evolved

through gene duplication. These repeats act as a structural building block and

form more than one compact structural and functional unit called a repeat domain.

The protein families comprising similar structural repeats are mainly involved in

protein-protein interactions as well as binding to other ligand molecules. The

identification of internal sequence repeats in the primary structure is not

sufficient for the analysis of structural repeats. Thus, a new method called

ProSTRIP has been developed using dynamic programming to find the similar

structural repeats in a three-dimensional protein structure. The detection of

these repeats is made by calculating the protein backbone Calpha angles. An

internet computing server is also created by implementing this method and enables

graphical visualization of the results. It can be freely accessed at

http://cluster.physics.iisc.ernet.in/prostrip/.

Copyright 2010 Elsevier Ltd. All rights reserved.

DOI: 10.1016/j.compbiolchem.2010.03.006

PMID: 20430700 [Indexed for MEDLINE]

2028. IEEE/ACM Trans Comput Biol Bioinform. 2010 Apr-Jun;7(2):309-22. doi:

10.1109/TCBB.2008.28.

Alignments of RNA structures.

Blin G(1), Denise A, Dulucq S, Herrbach C, Touzet H.

Author information:

(1)Institut Gaspard Monge, UMR 8049 CNRS, Université Paris-Est, 5 Bd

Descartes-Champs sur Marne, 77454 Marne La Vallée Cedex 2, France.

gblin@univ-mlv.fr

We describe a theoretical unifying framework to express the comparison of RNA

structures, which we call alignment hierarchy. This framework relies on the

definition of common supersequences for arc-annotated sequences and encompasses

the main existing models for RNA structure comparison based on trees and

arc-annotated sequences with a variety of edit operations. It also gives rise to

edit models that have not been studied yet. We provide a thorough analysis of the

alignment hierarchy, including a new polynomial-time algorithm and an

NP-completeness proof. The polynomial-time algorithm involves biologically

relevant edit operations such as pairing or unpairing nucleotides. It has been

implemented in a software, called gardenia, which is available at the Web server

http://bioinfo.lifl.fr/RNA/gardenia.

DOI: 10.1109/TCBB.2008.28

PMID: 20431150 [Indexed for MEDLINE]

2029. Nat Protoc. 2010 Apr;5(4):725-38. doi: 10.1038/nprot.2010.5. Epub 2010 Mar 25.

I-TASSER: a unified platform for automated protein structure and function

prediction.

Roy A(1), Kucukural A, Zhang Y.

Author information:

(1)Center for Computational Medicine and Bioinformatics, University of Michigan,

Ann Arbor, Michigan, USA.

The iterative threading assembly refinement (I-TASSER) server is an integrated

platform for automated protein structure and function prediction based on the

sequence-to-structure-to-function paradigm. Starting from an amino acid sequence,

I-TASSER first generates three-dimensional (3D) atomic models from multiple

threading alignments and iterative structural assembly simulations. The function

of the protein is then inferred by structurally matching the 3D models with other

known proteins. The output from a typical server run contains full-length

secondary and tertiary structure predictions, and functional annotations on

ligand-binding sites, Enzyme Commission numbers and Gene Ontology terms. An

estimate of accuracy of the predictions is provided based on the confidence score

of the modeling. This protocol provides new insights and guidelines for designing

of online server systems for the state-of-the-art protein structure and function

predictions. The server is available at

http://zhanglab.ccmb.med.umich.edu/I-TASSER.

DOI: 10.1038/nprot.2010.5

PMCID: PMC2849174

PMID: 20360767 [Indexed for MEDLINE]

2030. Nucleic Acids Res. 2010 Apr;38(7):e103. doi: 10.1093/nar/gkq021. Epub 2010 Jan

31.

DotKnot: pseudoknot prediction using the probability dot plot under a refined

energy model.

Sperschneider J(1), Datta A.

Author information:

(1)School of Computer Science and Software Engineering, The University of Western

Australia, Perth, WA 6009, Australia. janaspe@csse.uwa.edu.au

RNA pseudoknots are functional structure elements with key roles in viral and

cellular processes. Prediction of a pseudoknotted minimum free energy structure

is an NP-complete problem. Practical algorithms for RNA structure prediction

including restricted classes of pseudoknots suffer from high runtime and poor

accuracy for longer sequences. A heuristic approach is to search for promising

pseudoknot candidates in a sequence and verify those. Afterwards, the detected

pseudoknots can be further analysed using bioinformatics or laboratory

techniques. We present a novel pseudoknot detection method called DotKnot that

extracts stem regions from the secondary structure probability dot plot and

assembles pseudoknot candidates in a constructive fashion. We evaluate pseudoknot

free energies using novel parameters, which have recently become available. We

show that the conventional probability dot plot makes a wide class of pseudoknots

including those with bulged stems manageable in an explicit fashion. The energy

parameters now become the limiting factor in pseudoknot prediction. DotKnot is an

efficient method for long sequences, which finds pseudoknots with higher accuracy

compared to other known prediction algorithms. DotKnot is accessible as a web

server at http://dotknot.csse.uwa.edu.au.

DOI: 10.1093/nar/gkq021

PMCID: PMC2853144

PMID: 20123730 [Indexed for MEDLINE]

2031. Peptides. 2010 Apr;31(4):574-8. doi: 10.1016/j.peptides.2009.12.026. Epub 2010

Jan 4.

SecretP: a new method for predicting mammalian secreted proteins.

Yu L(1), Guo Y, Zhang Z, Li Y, Li M, Li G, Xiong W, Zeng Y.

Author information:

(1)College of Chemistry, Sichuan University, Chengdu 610064, PR China.

In contrast to a large number of classically secreted proteins (CSPs) and

non-secreted proteins (NSPs), only a few proteins have been experimentally proved

to enter non-classical secretory pathways. So it is difficult to identify

non-classically secreted proteins (NCSPs), and no methods are available for

distinguishing the three types of proteins simultaneously. In order to solve this

problem, a data mining has been taken firstly, and mammalian proteins exported

via ER-Golgi-independent pathways are collected through extensive literature

searches. In this paper, a support vector machine (SVM)-based ternary classifier

named SecretP is proposed to predict mammalian secreted proteins by using

pseudo-amino acid composition (PseAA) and five additional features. When

distinguishing the three types of proteins, SecretP yielded an accuracy of

88.79%. Evaluating the performance of our method by an independent test set of 92

human proteins, 76 of them are correctly predicted as NCSPs. When performed on

another public independent data set, the prediction result of SecretP is

comparable to those of other existing computational methods. Therefore, SecretP

can be a useful supplementary tool for future secretome studies. The web server

SecretP and all supplementary tables listed in this paper are freely available at

http://cic.scu.edu.cn/bioinformatics/secretp/index.htm.

Copyright (c) 2010 Elsevier Inc. All rights reserved.

DOI: 10.1016/j.peptides.2009.12.026

PMID: 20045033 [Indexed for MEDLINE]

2032. PLoS One. 2010 Apr 1;5(4):e9931. doi: 10.1371/journal.pone.0009931.

A new method for predicting the subcellular localization of eukaryotic proteins

with both single and multiple sites: Euk-mPLoc 2.0.

Chou KC(1), Shen HB.

Author information:

(1)Gordon Life Science Institute, San Diego, California, USA.

kcchou@gordonlifescience.org

Information of subcellular locations of proteins is important for in-depth

studies of cell biology. It is very useful for proteomics, system biology and

drug development as well. However, most existing methods for predicting protein

subcellular location can only cover 5 to 12 location sites. Also, they are

limited to deal with single-location proteins and hence failed to work for

multiplex proteins, which can simultaneously exist at, or move between, two or

more location sites. Actually, multiplex proteins of this kind usually posses

some important biological functions worthy of our special notice. A new predictor

called "Euk-mPLoc 2.0" is developed by hybridizing the gene ontology information,

functional domain information, and sequential evolutionary information through

three different modes of pseudo amino acid composition. It can be used to

identify eukaryotic proteins among the following 22 locations: (1) acrosome, (2)

cell wall, (3) centriole, (4) chloroplast, (5) cyanelle, (6) cytoplasm, (7)

cytoskeleton, (8) endoplasmic reticulum, (9) endosome, (10) extracell, (11) Golgi

apparatus, (12) hydrogenosome, (13) lysosome, (14) melanosome, (15) microsome

(16) mitochondria, (17) nucleus, (18) peroxisome, (19) plasma membrane, (20)

plastid, (21) spindle pole body, and (22) vacuole. Compared with the existing

methods for predicting eukaryotic protein subcellular localization, the new

predictor is much more powerful and flexible, particularly in dealing with

proteins with multiple locations and proteins without available accession

numbers. For a newly-constructed stringent benchmark dataset which contains both

single- and multiple-location proteins and in which none of proteins has pairwise

sequence identity to any other in a same location, the overall jackknife success

rate achieved by Euk-mPLoc 2.0 is more than 24% higher than those by any of the

existing predictors. As a user-friendly web-server, Euk-mPLoc 2.0 is freely

accessible at http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/. For a query

protein sequence of 400 amino acids, it will take about 15 seconds for the

web-server to yield the predicted result; the longer the sequence is, the more

time it may usually need. It is anticipated that the novel approach and the

powerful predictor as presented in this paper will have a significant impact to

Molecular Cell Biology, System Biology, Proteomics, Bioinformatics, and Drug

Development.

DOI: 10.1371/journal.pone.0009931

PMCID: PMC2848569

PMID: 20368981 [Indexed for MEDLINE]

2033. Protein Pept Lett. 2010 Apr;17(4):464-72.

Predicting protein subcellular locations with feature selection and analysis.

Cai Y(1), He J, Li X, Feng K, Lu L, Feng K, Kong X, Lu W.

Author information:

(1)State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiao Tong

University, 197 Rui Jin Road II, Shanghai 200025, China.

caiyudong@staff.shu.edu.cn

In this paper, we propose a strategy to predict the subcellular locations of

proteins by combining various feature selection methods. Firstly, proteins are

coded by amino-acid composition and physicochemical properties, then these

features are arranged by Minimum Redundancy Maximum Relevance method and further

filtered by feature selection procedure. Nearest Neighbor Algorithm is used as a

prediction model to predict the protein subcellular locations, and gains a

correct prediction rate of 70.63%, evaluated by Jackknife cross-validation.

Results of feature selection also enable us to identify the most important

protein properties. The prediction software is available for public access on the

website http://chemdata.shu.edu.cn/sub22/, which may play a important

complementary role to a series of web-server predictors summarized recently in a

review by Chou and Shen (Chou, K.C., Shen, H.B. Natural Science, 2009, 2, 63-92,

http://www.scirp.org/journal/NS/).

PMID: 19995336 [Indexed for MEDLINE]

2034. Proteins. 2010 Apr;78(5):1195-211. doi: 10.1002/prot.22639.

Detection of multiscale pockets on protein surfaces using mathematical

morphology.

Kawabata T(1).

Author information:

(1)Graduate School of Information Science, Nara Institute of Science and

Technology, Takayama 8916-5, Ikoma, Nara 630-0192, Japan. takawaba@is.naist.jp

Detection of pockets on protein surfaces is an important step toward finding the

binding sites of small molecules. In a previous study, we defined a pocket as a

space into which a small spherical probe can enter, but a large probe cannot. The

radius of the large probes corresponds to the shallowness of pockets. We showed

that each type of binding molecule has a characteristic shallowness distribution.

In this study, we introduced fundamental changes to our previous algorithm by

using a 3D grid representation of proteins and probes, and the theory of

mathematical morphology. We invented an efficient algorithm for calculating deep

and shallow pockets (multiscale pockets) simultaneously, using several different

sizes of spherical probes (multiscale probes). We implemented our algorithm as a

new program, ghecom (grid-based HECOMi finder). The statistics of calculated

pockets for the structural dataset showed that our program had a higher

performance of detecting binding pockets, than four other popular pocket-finding

programs proposed previously. The ghecom also calculates the shallowness of

binding ligands, R(inaccess) (minimum radius of inaccessible spherical probes)

that can be obtained from the multiscale molecular volume. We showed that each

part of the binding molecule had a bias toward a specific range of shallowness.

These findings will be useful for predicting the types of molecules that will be

most likely to bind putative binding pockets, as well as the configurations of

binding molecules. The program ghecom is available through the Web server

(http://biunit.naist.jp/ghecom).

DOI: 10.1002/prot.22639

PMID: 19938154 [Indexed for MEDLINE]

2035. BMC Res Notes. 2010 Mar 31;3:87. doi: 10.1186/1756-0500-3-87.

RCDI/eRCDI: a web-server to estimate codon usage deoptimization.

Puigbò P(1), Aragonès L, Garcia-Vallvé S.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, Bethesda, MD 20894, USA.

puigboap@ncbi.nlm.nih.gov.

BACKGROUND: The Relative Codon Deoptimization Index (RCDI) was developed by

Mueller et al. (2006) as measure of codon deoptimization by comparing how similar

is the codon usage of a gene and the codon usage of a reference genome.

FINDINGS: RCDI/eRCDI is a web application server that calculates the Relative

Codon Deoptimization Index and a new expected value for the RCDI (eRCDI). The

RCDI is used to estimate the similarity of the codon frequencies of a specific

gene in comparison to a given reference genome. The eRCDI is determined by

generating random sequences with similar G+C and amino acid composition to the

input sequences and may be used as an indicator of the significance of the RCDI

values. RCDI/eRCDI is freely available at http://genomes.urv.cat/CAIcal/RCDI.

CONCLUSIONS: This web server will be a useful tool for genome analysis, to

understand host-virus phylogenetic relationships or to infer the potential host

range of a virus and its replication strategy, as well as in experimental

virology to ease the step of gene design for heterologous protein expression.

DOI: 10.1186/1756-0500-3-87

PMCID: PMC2853550

PMID: 20356391

2036. BMC Bioinformatics. 2010 Mar 30;11:161. doi: 10.1186/1471-2105-11-161.

ProbFAST: Probabilistic functional analysis system tool.

Silva IT(1), Vêncio RZ, Oliveira TY, Molfetta GA, Silva WA Jr.

Author information:

(1)Department of Genetics, Faculty of Medicine, University of São Paulo, Ribeirão

Preto, Brazil. itojal@usp.br

BACKGROUND: The post-genomic era has brought new challenges regarding the

understanding of the organization and function of the human genome. Many of these

challenges are centered on the meaning of differential gene regulation under

distinct biological conditions and can be performed by analyzing the Multiple

Differential Expression (MDE) of genes associated with normal and abnormal

biological processes. Currently MDE analyses are limited to usual methods of

differential expression initially designed for paired analysis.

RESULTS: We proposed a web platform named ProbFAST for MDE analysis which uses

Bayesian inference to identify key genes that are intuitively prioritized by

means of probabilities. A simulated study revealed that our method gives a better

performance when compared to other approaches and when applied to public

expression data, we demonstrated its flexibility to obtain relevant genes

biologically associated with normal and abnormal biological processes.

CONCLUSIONS: ProbFAST is a free accessible web-based application that enables MDE

analysis on a global scale. It offers an efficient methodological approach for

MDE analysis of a set of genes that are turned on and off related to functional

information during the evolution of a tumor or tissue differentiation. ProbFAST

server can be accessed at http://gdm.fmrp.usp.br/probfast.

DOI: 10.1186/1471-2105-11-161

PMCID: PMC2868004

PMID: 20353576 [Indexed for MEDLINE]

2037. BMC Bioinformatics. 2010 Mar 30;11:160. doi: 10.1186/1471-2105-11-160.

Identification of NAD interacting residues in proteins.

Ansari HR(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector 39A, Chandigarh, 160036, India.

BACKGROUND: Small molecular cofactors or ligands play a crucial role in the

proper functioning of cells. Accurate annotation of their target proteins and

binding sites is required for the complete understanding of reaction mechanisms.

Nicotinamide adenine dinucleotide (NAD+ or NAD) is one of the most commonly used

organic cofactors in living cells, which plays a critical role in cellular

metabolism, storage and regulatory processes. In the past, several NAD binding

proteins (NADBP) have been reported in the literature, which are responsible for

a wide-range of activities in the cell. Attempts have been made to derive a rule

for the binding of NAD+ to its target proteins. However, so far an efficient

model could not be derived due to the time consuming process of structure

determination, and limitations of similarity based approaches. Thus a sequence

and non-similarity based method is needed to characterize the NAD binding sites

to help in the annotation. In this study attempts have been made to predict NAD

binding proteins and their interacting residues (NIRs) from amino acid sequence

using bioinformatics tools.

RESULTS: We extracted 1556 proteins chains from 555 NAD binding proteins whose

structure is available in Protein Data Bank. Then we removed all redundant

protein chains and finally obtained 195 non-redundant NAD binding protein chains,

where no two chains have more than 40% sequence identity. In this study all

models were developed and evaluated using five-fold cross validation technique on

the above dataset of 195 NAD binding proteins. While certain type of residues are

preferred (e.g. Gly, Tyr, Thr, His) in NAD interaction, residues like Ala, Glu,

Leu, Lys are not preferred. A support vector machine (SVM) based method has been

developed using various window lengths of amino acid sequence for predicting NAD

interacting residues and obtained maximum Matthew's correlation coefficient (MCC)

0.47 with accuracy 74.13% at window length 17. We also developed a SVM based

method using evolutionary information in the form of position specific scoring

matrix (PSSM) and obtained maximum MCC 0.75 with accuracy 87.25%.

CONCLUSION: For the first time a sequence-based method has been developed for the

prediction of NAD binding proteins and their interacting residues, in the absence

of any prior structural information. The present model will aid in the

understanding of NAD+ dependent mechanisms of action in the cell. To provide

service to the scientific community, we have developed a user-friendly web

server, which is available from URL http://www.imtech.res.in/raghava/nadbinder/.

DOI: 10.1186/1471-2105-11-160

PMCID: PMC2853471

PMID: 20353553 [Indexed for MEDLINE]

2038. BMC Bioinformatics. 2010 Mar 17;11:138. doi: 10.1186/1471-2105-11-138.

An optimized TOPS+ comparison method for enhanced TOPS models.

Veeramalai M(1), Gilbert D, Valiente G.

Author information:

(1)Joint Center for Molecular Modeling, Sanford-Burnham Medical Research

Institute, La Jolla, CA 92037, USA. mallikav@burnham.org

BACKGROUND: Although methods based on highly abstract descriptions of protein

structures, such as VAST and TOPS, can perform very fast protein structure

comparison, the results can lack a high degree of biological significance.

Previously we have discussed the basic mechanisms of our novel method for

structure comparison based on our TOPS+ model (Topological descriptions of

Protein Structures Enhanced with Ligand Information). In this paper we show how

these results can be significantly improved using parameter optimization, and we

call the resulting optimised TOPS+ method as advanced TOPS+ comparison method

i.e. advTOPS+.

RESULTS: We have developed a TOPS+ string model as an improvement to the TOPS 123

graph model by considering loops as secondary structure elements (SSEs) in

addition to helices and strands, representing ligands as first class objects, and

describing interactions between SSEs, and SSEs and ligands, by incoming and

outgoing arcs, annotating SSEs with the interaction direction and type.

Benchmarking results of an all-against-all pairwise comparison using a large

dataset of 2,620 non-redundant structures from the PDB40 dataset 4 demonstrate

the biological significance, in terms of SCOP classification at the superfamily

level, of our TOPS+ comparison method.

CONCLUSIONS: Our advanced TOPS+ comparison shows better performance on the PDB40

dataset 4 compared to our basic TOPS+ method, giving 90% accuracy for SCOP

alpha+beta; a 6% increase in accuracy compared to the TOPS and basic TOPS+

methods. It also outperforms the TOPS, basic TOPS+ and SSAP comparison methods on

the Chew-Kedem dataset 5, achieving 98% accuracy.

SOFTWARE AVAILABILITY: The TOPS+ comparison server is available at

http://balabio.dcs.gla.ac.uk/mallika/WebTOPS/.

DOI: 10.1186/1471-2105-11-138

PMCID: PMC2858036

PMID: 20236520 [Indexed for MEDLINE]

2039. BMC Bioinformatics. 2010 Mar 16;11:133. doi: 10.1186/1471-2105-11-133.

MapMi: automated mapping of microRNA loci.

Guerra-Assunção JA(1), Enright AJ.

Author information:

(1)European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton,

Cambridge CB10 1SD, UK.

BACKGROUND: A large effort to discover microRNAs (miRNAs) has been under way.

Currently miRBase is their primary repository, providing annotations of primary

sequences, precursors and probable genomic loci. In many cases miRNAs are

identical or very similar between related (or in some cases more distant)

species. However, miRBase focuses on those species for which miRNAs have been

directly confirmed. Secondly, specific miRNAs or their loci are sometimes not

annotated even in well-covered species. We sought to address this problem by

developing a computational system for automated mapping of miRNAs within and

across species. Given the sequence of a known miRNA in one species it is

relatively straightforward to determine likely loci of that miRNA in other

species. Our primary goal is not the discovery of novel miRNAs but the mapping of

validated miRNAs in one species to their most likely orthologues in other

species.

RESULTS: We present MapMi, a computational system for automated miRNA mapping

across and within species. This method has a sensitivity of 92.20% and a

specificity of 97.73%. Using the latest release (v14) of miRBase, we obtained

10,944 unannotated potential miRNAs when MapMi was applied to all 21 species in

Ensembl Metazoa release 2 and 46 species from Ensembl release 55.

CONCLUSIONS: The pipeline and an associated web-server for mapping miRNAs are

freely available on http://www.ebi.ac.uk/enright-srv/MapMi/. In addition

precomputed miRNA mappings of miRBase miRNAs across a large number of species are

provided.

DOI: 10.1186/1471-2105-11-133

PMCID: PMC2858034

PMID: 20233390 [Indexed for MEDLINE]

2040. Bioinformatics. 2010 Mar 15;26(6):814-21. doi: 10.1093/bioinformatics/btq024.

Epub 2010 Jan 26.

Predicting biodegradation products and pathways: a hybrid knowledge- and machine

learning-based approach.

Wicker J(1), Fenner K, Ellis L, Wackett L, Kramer S.

Author information:

(1)Institut für Informatik/I12, Technische Universität München, D-85748 Garching

b. München, Germany.

MOTIVATION: Current methods for the prediction of biodegradation products and

pathways of organic environmental pollutants either do not take into account

domain knowledge or do not provide probability estimates. In this article, we

propose a hybrid knowledge- and machine learning-based approach to overcome these

limitations in the context of the University of Minnesota Pathway Prediction

System (UM-PPS). The proposed solution performs relative reasoning in a machine

learning framework, and obtains one probability estimate for each

biotransformation rule of the system. As the application of a rule then depends

on a threshold for the probability estimate, the trade-off between recall

(sensitivity) and precision (selectivity) can be addressed and leveraged in

practice.

RESULTS: Results from leave-one-out cross-validation show that a recall and

precision of approximately 0.8 can be achieved for a subset of 13 transformation

rules. Therefore, it is possible to optimize precision without compromising

recall. We are currently integrating the results into an experimental version of

the UM-PPS server.

AVAILABILITY: The program is freely available on the web at

http://wwwkramer.in.tum.de/research/applications/biodegradation/data.

CONTACT: kramer@in.tum.de.

DOI: 10.1093/bioinformatics/btq024

PMID: 20106820 [Indexed for MEDLINE]

2041. Bioinformatics. 2010 Mar 15;26(6):845-6. doi: 10.1093/bioinformatics/btq030. Epub

2010 Jan 26.

FineStr: a web server for single-base-resolution nucleosome positioning.

Gabdank I(1), Barash D, Trifonov EN.

Author information:

(1)Department of Computer Science, Ben Gurion University of the Negev, Beer Sheva

84105, Israel. gabdank@cs.bgu.ac.il

SUMMARY: The DNA in eukaryotic cells is packed into the chromatin that is

composed of nucleosomes. Positioning of the nucleosome core particles on the

sequence is a problem of great interest because of the role nucleosomes play in

different cellular processes including gene regulation. Using the sequence

structure of 10.4 base DNA repeat presented in our previous works and nucleosome

core DNA sequences database, we have derived the complete nucleosome DNA

bendability matrix of Caenorhabditis elegans. We have developed a web server

named FineStr that allows users to upload genomic sequences in FASTA format and

to perform a single-base-resolution nucleosome mapping on them.

AVAILABILITY: FineStr server is freely available for use on the web at

http:/www.cs.bgu.ac.il/ approximately nucleom. The site contains a help file with

explanation regarding the exact usage.

CONTACT: gabdank@cs.bgu.ac.il.

DOI: 10.1093/bioinformatics/btq030

PMID: 20106816 [Indexed for MEDLINE]

2042. PLoS One. 2010 Mar 15;5(3):e9695. doi: 10.1371/journal.pone.0009695.

FaaPred: a SVM-based prediction method for fungal adhesins and adhesin-like

proteins.

Ramana J(1), Gupta D.

Author information:

(1)Structural and Computational Biology Group, International Centre for Genetic

Engineering and Biotechnology, Aruna Asaf Ali Marg, New Delhi, India.

Adhesion constitutes one of the initial stages of infection in microbial diseases

and is mediated by adhesins. Hence, identification and comprehensive knowledge of

adhesins and adhesin-like proteins is essential to understand adhesin mediated

pathogenesis and how to exploit its therapeutic potential. However, the knowledge

about fungal adhesins is rudimentary compared to that of bacterial adhesins. In

addition to host cell attachment and mating, the fungal adhesins play a

significant role in homotypic and xenotypic aggregation, foraging and biofilm

formation. Experimental identification of fungal adhesins is labor- as well as

time-intensive. In this work, we present a Support Vector Machine (SVM) based

method for the prediction of fungal adhesins and adhesin-like proteins. The SVM

models were trained with different compositional features, namely, amino acid,

dipeptide, multiplet fractions, charge and hydrophobic compositions, as well as

PSI-BLAST derived PSSM matrices. The best classifiers are based on compositional

properties as well as PSSM and yield an overall accuracy of 86%. The prediction

method based on best classifiers is freely accessible as a world wide web based

server at http://bioinfo.icgeb.res.in/faap. This work will aid rapid and rational

identification of fungal adhesins, expedite the pace of experimental

characterization of novel fungal adhesins and enhance our knowledge about role of

adhesins in fungal infections.

DOI: 10.1371/journal.pone.0009695

PMCID: PMC2837750

PMID: 20300572 [Indexed for MEDLINE]

2043. Bioinformatics. 2010 Mar 1;26(5):692-3. doi: 10.1093/bioinformatics/btq019. Epub

2010 Jan 19.

iDBPs: a web server for the identification of DNA binding proteins.

Nimrod G(1), Schushan M, Szilágyi A, Leslie C, Ben-Tal N.

Author information:

(1)Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel

Aviv University, Ramat Aviv 69978, Israel.

SUMMARY: The iDBPs server uses the three-dimensional (3D) structure of a query

protein to predict whether it binds DNA. First, the algorithm predicts the

functional region of the protein based on its evolutionary profile; the

assumption is that large clusters of conserved residues are good markers of

functional regions. Next, various characteristics of the predicted functional

region as well as global features of the protein are calculated, such as the

average surface electrostatic potential, the dipole moment and cluster-based

amino acid conservation patterns. Finally, a random forests classifier is used to

predict whether the query protein is likely to bind DNA and to estimate the

prediction confidence. We have trained and tested the classifier on various

datasets and shown that it outperformed related methods. On a dataset that

reflects the fraction of DNA binding proteins (DBPs) in a proteome, the area

under the ROC curve was 0.90. The application of the server to an updated version

of the N-Func database, which contains proteins of unknown function with solved

3D-structure, suggested new putative DBPs for experimental studies.

AVAILABILITY: http://idbps.tau.ac.il/

DOI: 10.1093/bioinformatics/btq019

PMCID: PMC2828122

PMID: 20089514 [Indexed for MEDLINE]

2044. Bioinformatics. 2010 Mar 1;26(5):680-2. doi: 10.1093/bioinformatics/btq003. Epub

2010 Jan 6.

CD-HIT Suite: a web server for clustering and comparing biological sequences.

Huang Y(1), Niu B, Gao Y, Fu L, Li W.

Author information:

(1)California Institute for Telecommunications and Information Technology,

University of California San Diego, La Jolla, CA, USA.

CD-HIT is a widely used program for clustering and comparing large biological

sequence datasets. In order to further assist the CD-HIT users, we significantly

improved this program with more functions and better accuracy, scalability and

flexibility. Most importantly, we developed a new web server, CD-HIT Suite, for

clustering a user-uploaded sequence dataset or comparing it to another dataset at

different identity levels. Users can now interactively explore the clusters

within web browsers. We also provide downloadable clusters for several public

databases (NCBI NR, Swissprot and PDB) at different identity levels.AVAILABILITY:

Free access at http://cd-hit.org

DOI: 10.1093/bioinformatics/btq003

PMCID: PMC2828112

PMID: 20053844 [Indexed for MEDLINE]

2045. Microbiology. 2010 Mar;156(Pt 3):849-59. doi: 10.1099/mic.0.035790-0. Epub 2009

Dec 3.

Connecting parts with processes: SubtiWiki and SubtiPathways integrate gene and

pathway annotation for Bacillus subtilis.

Lammers CR(1), Flórez LA, Schmeisky AG, Roppel SF, Mäder U, Hamoen L, Stülke J.

Author information:

(1)Department of General Microbiology, Institute of Microbiology and Genetics,

Georg-August University Göttingen, Germany.

Bacillus subtilis is the model organism for a large group of Gram-positive

bacteria, the Firmicutes. Several online databases have been established over

time to manage its genetic and metabolic information, but they differ greatly in

their rate of update and their focus on B. subtilis. Therefore, a European

systems biology consortium called for an integrated solution that empowers its

users to enrich online content. To meet this goal we created SubtiWiki and

SubtiPathways, two complementary online tools for gene and pathway information on

B. subtilis 168. SubtiWiki (http://subtiwiki.uni-goettingen.de/ ) is a scientific

wiki for all genes of B. subtilis and their protein or RNA products. Each gene

page contains a summary of the most important information; sections on the gene,

its product and expression; sections concerning biological materials and

laboratories; and a list of references. SubtiWiki has been seeded with key

content and can be extended by any researcher after a simple registration, thus

keeping it always up to date. As a complement, SubtiPathways

(http://subtipathways.uni-goettingen.de/) is an online tool for navigation of the

metabolism of B. subtilis and its regulation. Each SubtiPathways diagram presents

a metabolic pathway with its participating enzymes, together with the regulatory

mechanisms that act on their expression and activity, in an intuitive interface

that is based on Google Maps. Together, SubtiWiki and SubtiPathways provide an

integrated view of the processes that make up B. subtilis and its components,

making it the most comprehensive web resource for B. subtilis researchers.

DOI: 10.1099/mic.0.035790-0

PMID: 19959575 [Indexed for MEDLINE]

2046. Nucleic Acids Res. 2010 Mar;38(5):1711-22. doi: 10.1093/nar/gkp1054. Epub 2009

Dec 30.

Computing folding pathways between RNA secondary structures.

Dotu I(1), Lorenz WA, Van Hentenryck P, Clote P.

Author information:

(1)Department of Computer Science, Brown University, PO Box 1910 Providence, RI

02912, USA.

Given an RNA sequence and two designated secondary structures A, B, we describe a

new algorithm that computes a nearly optimal folding pathway from A to B. The

algorithm, RNAtabupath, employs a tabu semi-greedy heuristic, known to be an

effective search strategy in combinatorial optimization. Folding pathways,

sometimes called routes or trajectories, are computed by RNAtabupath in a

fraction of the time required by the barriers program of Vienna RNA Package. We

benchmark RNAtabupath with other algorithms to compute low energy folding

pathways between experimentally known structures of several conformational

switches. The RNApathfinder web server, source code for algorithms to compute and

analyze pathways and supplementary data are available at

http://bioinformatics.bc.edu/clotelab/RNApathfinder.

DOI: 10.1093/nar/gkp1054

PMCID: PMC2836545

PMID: 20044352 [Indexed for MEDLINE]

2047. Protein Pept Lett. 2010 Mar;17(3):287-96.

Improve the prediction of RNA-binding residues using structural neighbours.

Li Q(1), Cao Z, Liu H.

Author information:

(1)School of Life Sciences, Hefei National Laboratory for Physical Sciences at

Microscale, University of Science and Technology of China (USTC), Hefei, Anhui,

230027, China. jeele@mail.ustc.edu.cn

The interactions between RNA-binding proteins (RBPs) with RNA play key roles in

managing some of the cell's basic functions. The identification and prediction of

RNA binding sites is important for understanding the RNA-binding mechanism.

Computational approaches are being developed to predict RNA-binding residues

based on the sequence- or structure-derived features. To achieve higher

prediction accuracy, improvements on current prediction methods are necessary. We

identified that the structural neighbors of RNA-binding and non-RNA-binding

residues have different amino acid compositions. Combining this structure-derived

feature with evolutionary (PSSM) and other structural information (secondary

structure and solvent accessibility) significantly improves the predictions over

existing methods. Using a multiple linear regression approach and 6-fold cross

validation, our best model can achieve an overall correct rate of 87.8% and MCC

of 0.47, with a specificity of 93.4%, correctly predict 52.4% of the RNA-binding

residues for a dataset containing 107 non-homologous RNA-binding proteins.

Compared with existing methods, including the amino acid compositions of

structure neighbors lead to clearly improvement. A web server was developed for

predicting RNA binding residues in a protein sequence (or structure),which is

available at http://mcgill.3322.org/RNA/.

PMID: 19508202 [Indexed for MEDLINE]

2048. Wien Med Wochenschr. 2010 Mar;160(5-6):129-38. doi: 10.1007/s10354-009-0738-9.

Blood coagulation disorders in septic patients.

Knoebl P(1).

Author information:

(1)Division of Hematology and Hemostasis, Department of Medicine I, Medical

University of Vienna, Vienna, Austria. paul.knoebl@meduniwien.ac.at

Comment in

Wien Med Wochenschr. 2010 Mar;160(5-6):105-6.

Host defense and blood coagulation are tightly connected and interacting systems,

necessary for the integrity of an organism. Complex mechanisms regulate the

intensity of a host response to invading pathogens or other potentially dangerous

situations. Under regular conditions, this response is limited in time and

located to the site of injury. Sometimes, however, systemic host response is

overwhelming and disproportional and causes damage, not cure. Dependent on the

genetical predisposition of the host, its current immunocompetence, or the type

of injury, the reaction leads to the clinical picture of the different degrees of

sepsis. Septic organ dysfunction is caused by intravascular fibrin deposition as

a result of coagulation activation, anticoagulant breakdown, and shut down of

fibrinolysis. This article describes the major pathophysiologic reactions in

these situations and presents www.SepDIC.eu, an online tool on sepsis and

associated coagulopathy.

DOI: 10.1007/s10354-009-0738-9

PMID: 20364416 [Indexed for MEDLINE]

2049. Bioinformation. 2010 Feb 28;4(8):344-6.

The microarray manual curation tool (MMCT): A Web server for microarray probe

evaluations.

Tulpan D(1), Belliveau L, Léger S.

Author information:

(1)Institute of Information Technology National Research Council of Canada

Moncton, New Brunswick, E1A 7R1, Canada. dan.tulpan@nrc-cnrc.gc.ca

Quality control of probe sequences is a major concern in microarray technology.

The presence of poor quality probes has a negative impact on the microarray data

analysis process. The Microarray Manual Curation Tool (MMCT) is a web server

application that provides computational and visual means to investigate the

quality of individual probes for oligo microarrays. The MMCT quality metrics

assess the free energy of hybridization and the secondary structure of duplexes

formed by selected targets and probes, which are specific to various microarray

platforms.AVAILABILITY: http://www.nrcbioinformatics.ca/mmct.

PMCID: PMC2951669

PMID: 20975897

2050. Bioinformation. 2010 Feb 28;4(8):341-3.

WebFARM: web server for finite automated restriction mapping.

Singh TR(1).

Author information:

(1)Bioinformatics Centre, School of Biotechnology, DAVV, Indore, India.

tiratharaj@gmail.com

Restriction endonucleases are indispensable tools in molecular biology and

biotechnology. Type II restriction endonucleases are part of restriction

modification systems. DNA fragment extraction and restriction mapping are the

basis for several biotechnological activities. WebFARM is a server application

for identifying restriction endonuclease recognition sites and to give

information regarding restriction mapping for given nucleotide sequences. WebFARM

analyses given nucleotide sequence and identify restriction site for selected

restriction endonucleases. It will also provide frequency of restriction for each

restriction endonuclease.AVAILABILITY: http://webfarm.bioinfoindia.org/

PMCID: PMC2951673

PMID: 20975896

2051. BMC Bioinformatics. 2010 Feb 24;11:102. doi: 10.1186/1471-2105-11-102.

Reconstructing genome trees of prokaryotes using overlapping genes.

Cheng CH(1), Yang CH, Chiu HT, Lu CL.

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu 300, Taiwan.

BACKGROUND: Overlapping genes (OGs) are defined as adjacent genes whose coding

sequences overlap partially or entirely. In fact, they are ubiquitous in

microbial genomes and more conserved between species than non-overlapping genes.

Based on this property, we have previously implemented a web server, named

OGtree, that allows the user to reconstruct genome trees of some prokaryotes

according to their pairwise OG distances. By analogy to the analyses of gene

content and gene order, the OG distance between two genomes we defined was based

on a measure of combining OG content (i.e., the normalized number of shared

orthologous OG pairs) and OG order (i.e., the normalized OG breakpoint distance)

in their whole genomes. A shortcoming of using the concept of breakpoints to

define the OG distance is its inability to analyze the OG distance of

multi-chromosomal genomes. In addition, the amount of overlapping coding

sequences between some distantly related prokaryotic genomes may be limited so

that it is hard to find enough OGs to properly evaluate their pairwise OG

distances.

RESULTS: In this study, we therefore define a new OG order distance that is based

on more biologically accurate rearrangements (e.g., reversals, transpositions and

translocations) rather than breakpoints and that is applicable to both

uni-chromosomal and multi-chromosomal genomes. In addition, we expand the term

"gene" to include both its coding sequence and regulatory regions so that two

adjacent genes whose coding sequences or regulatory regions overlap with each

other are considered as a pair of overlapping genes. This is because overlapping

of regulatory regions of distinct genes suggests that the regulation of

expression for these genes should be more or less interrelated. Based on these

modifications, we have reimplemented our OGtree as a new web server, named

OGtree2, and have also evaluated its accuracy of genome tree reconstruction on a

testing dataset consisting of 21 Proteobacteria genomes. Our experimental results

have finally shown that our current OGtree2 indeed outperforms its previous

version OGtree, as well as another similar server, called BPhyOG, significantly

in the quality of genome tree reconstruction, because the phylogenetic tree

obtained by OGtree2 is greatly congruent with the reference tree that coincides

with the taxonomy accepted by biologists for these Proteobacteria.

CONCLUSIONS: In this study, we have introduced a new web server OGtree2 at

http://bioalgorithm.life.nctu.edu.tw/OGtree2.0/ that can serve as a useful tool

for reconstructing more precise and robust genome trees of prokaryotes according

to their overlapping genes.

DOI: 10.1186/1471-2105-11-102

PMCID: PMC2845580

PMID: 20181237 [Indexed for MEDLINE]

2052. BMC Bioinformatics. 2010 Feb 17;11:91. doi: 10.1186/1471-2105-11-91.

BisoGenet: a new tool for gene network building, visualization and analysis.

Martin A(1), Ochagavia ME, Rabasa LC, Miranda J, Fernandez-de-Cossio J, Bringas

R.

Author information:

(1)Center for Genetic Engineering and Biotechnology, Havana, Cuba.

BACKGROUND: The increasing availability and diversity of omics data in the

post-genomic era offers new perspectives in most areas of biomedical research.

Graph-based biological networks models capture the topology of the functional

relationships between molecular entities such as gene, protein and small

compounds and provide a suitable framework for integrating and analyzing

omics-data. The development of software tools capable of integrating data from

different sources and to provide flexible methods to reconstruct, represent and

analyze topological networks is an active field of research in bioinformatics.

RESULTS: BisoGenet is a multi-tier application for visualization and analysis of

biomolecular relationships. The system consists of three tiers. In the data tier,

an in-house database stores genomics information, protein-protein interactions,

protein-DNA interactions, gene ontology and metabolic pathways. In the middle

tier, a global network is created at server startup, representing the whole data

on bioentities and their relationships retrieved from the database. The client

tier is a Cytoscape plugin, which manages user input, communication with the Web

Service, visualization and analysis of the resulting network.

CONCLUSION: BisoGenet is able to build and visualize biological networks in a

fast and user-friendly manner. A feature of Bisogenet is the possibility to

include coding relations to distinguish between genes and their products. This

feature could be instrumental to achieve a finer grain representation of the

bioentities and their relationships. The client application includes network

analysis tools and interactive network expansion capabilities. In addition, an

option is provided to allow other networks to be converted to BisoGenet. This

feature facilitates the integration of our software with other tools available in

the Cytoscape platform. BisoGenet is available at

http://bio.cigb.edu.cu/bisogenet-cytoscape/.

DOI: 10.1186/1471-2105-11-91

PMCID: PMC3098113

PMID: 20163717 [Indexed for MEDLINE]

2053. BMC Bioinformatics. 2010 Feb 17;11:89. doi: 10.1186/1471-2105-11-89.

Detection of distant evolutionary relationships between protein families using

theory of sequence profile-profile comparison.

Margelevicius M(1), Venclovas C.

Author information:

(1)Institute of Biotechnology, Graiciūno 8, LT-02241 Vilnius, Lithuania.

BACKGROUND: Detection of common evolutionary origin (homology) is a primary means

of inferring protein structure and function. At present, comparison of protein

families represented as sequence profiles is arguably the most effective homology

detection strategy. However, finding the best way to represent evolutionary

information of a protein sequence family in the profile, to compare profiles and

to estimate the biological significance of such comparisons, remains an active

area of research.

RESULTS: Here, we present a new homology detection method based on sequence

profile-profile comparison. The method has a number of new features including

position-dependent gap penalties and a global score system. Position-dependent

gap penalties provide a more biologically relevant way to represent and align

protein families as sequence profiles. The global score system enables an

analytical solution of the statistical parameters needed to estimate the

statistical significance of profile-profile similarities. The new method,

together with other state-of-the-art profile-based methods (HHsearch, COMPASS and

PSI-BLAST), is benchmarked in all-against-all comparison of a challenging set of

SCOP domains that share at most 20% sequence identity. For benchmarking, we use a

reference ("gold standard") free model-based evaluation framework. Evaluation

results show that at the level of protein domains our method compares favorably

to all other tested methods. We also provide examples of the new method

outperforming structure-based similarity detection and alignment. The

implementation of the new method both as a standalone software package and as a

web server is available at http://www.ibt.lt/bioinformatics/coma.

CONCLUSION: Due to a number of developments, the new profile-profile comparison

method shows an improved ability to match distantly related protein domains.

Therefore, the method should be useful for annotation and homology modeling of

uncharacterized proteins.

DOI: 10.1186/1471-2105-11-89

PMCID: PMC2837030

PMID: 20158924 [Indexed for MEDLINE]

2054. BMC Bioinformatics. 2010 Feb 12;11:87. doi: 10.1186/1471-2105-11-87.

Lists2Networks: integrated analysis of gene/protein lists.

Lachmann A(1), Ma'ayan A.

Author information:

(1)Department of Pharmacology and Systems Therapeutics, Systems Biology Center

New York, Mount Sinai School of Medicine, 1425 Madison Avenue, New York, NY

10029, USA.

BACKGROUND: Systems biologists are faced with the difficulty of analyzing results

from large-scale studies that profile the activity of many genes, RNAs and

proteins, applied in different experiments, under different conditions, and

reported in different publications. To address this challenge it is desirable to

compare the results from different related studies such as mRNA expression

microarrays, genome-wide ChIP-X, RNAi screens, proteomics and phosphoproteomics

experiments in a coherent global framework. In addition, linking high-content

multilayered experimental results with prior biological knowledge can be useful

for identifying functional themes and form novel hypotheses.

RESULTS: We present Lists2Networks, a web-based system that allows users to

upload lists of mammalian genes/proteins onto a server-based program for

integrated analysis. The system includes web-based tools to manipulate lists with

different set operations, to expand lists using existing mammalian networks of

protein-protein interactions, co-expression correlation, or background knowledge

co-annotation correlation, as well as to apply gene-list enrichment analyses

against many gene-list libraries of prior biological knowledge such as pathways,

gene ontology terms, kinase-substrate, microRNA-mRAN, and protein-protein

interactions, metabolites, and protein domains. Such analyses can be applied to

several lists at once against many prior knowledge libraries of gene-lists

associated with specific annotations. The system also contains features that

allow users to export networks and share lists with other users of the system.

CONCLUSIONS: Lists2Networks is a user friendly web-based software system expected

to significantly ease the computational analysis process for experimental systems

biologists employing high-throughput experiments at multiple layers of

regulation. The system is freely available at http://www.lists2networks.org.

DOI: 10.1186/1471-2105-11-87

PMCID: PMC2843617

PMID: 20152038 [Indexed for MEDLINE]

2055. PLoS One. 2010 Feb 11;5(2):e9171. doi: 10.1371/journal.pone.0009171.

The DIANA-mirExTra web server: from gene expression data to microRNA function.

Alexiou P(1), Maragkakis M, Papadopoulos GL, Simmosis VA, Zhang L, Hatzigeorgiou

AG.

Author information:

(1)Biomedical Sciences Research Center "Alexander Fleming", Institute of

Molecular Oncology, Varkiza, Greece.

BACKGROUND: High-throughput gene expression experiments are widely used to

identify the role of genes involved in biological conditions of interest.

MicroRNAs (miRNA) are regulatory molecules that have been functionally associated

with several developmental programs and their deregulation with diverse diseases

including cancer.

METHODOLOGY/PRINCIPAL FINDINGS: Although miRNA expression levels may not be

routinely measured in high-throughput experiments, a possible involvement of

miRNAs in the deregulation of gene expression can be computationally predicted

and quantified through analysis of overrepresented motifs in the deregulated

genes 3' untranslated region (3'UTR) sequences. Here, we introduce a

user-friendly web-server, DIANA-mirExTra (www.microrna.gr/mirextra) that allows

the comparison of frequencies of miRNA associated motifs between sets of genes

that can lead to the identification of miRNAs responsible for the deregulation of

large numbers of genes. To this end, we have investigated different approaches

and measures, and have practically implemented them on experimental data.

CONCLUSIONS/SIGNIFICANCE: On several datasets of miRNA overexpression and

repression experiments, our proposed approaches have successfully identified the

deregulated miRNA. Beyond the prediction of miRNAs responsible for the

deregulation of transcripts, the web-server provides extensive links to

DIANA-mirPath, a functional analysis tool incorporating miRNA targets in

biological pathways. Additionally, in case information about miRNA expression

changes is provided, the results can be filtered to display the analysis for

miRNAs of interest only.

DOI: 10.1371/journal.pone.0009171

PMCID: PMC2820085

PMID: 20161787 [Indexed for MEDLINE]

2056. J Proteome Res. 2010 Feb 5;9(2):1182-90. doi: 10.1021/pr900827b.

Trypano-PPI: a web server for prediction of unique targets in trypanosome

proteome by using electrostatic parameters of protein-protein interactions.

Rodriguez-Soca Y(1), Munteanu CR, Dorado J, Pazos A, Prado-Prado FJ,

González-Díaz H.

Author information:

(1)Department of Microbiology & Parasitology, Faculty of Pharmacy, University of

Santiago de Compostela, 15782, Santiago de Compostela, Spain.

Trypanosoma brucei causes African trypanosomiasis in humans (HAT or African

sleeping sickness) and Nagana in cattle. The disease threatens over 60 million

people and uncounted numbers of cattle in 36 countries of sub-Saharan Africa and

has a devastating impact on human health and the economy. On the other hand,

Trypanosoma cruzi is responsible in South America for Chagas disease, which can

cause acute illness and death, especially in young children. In this context, the

discovery of novel drug targets in Trypanosome proteome is a major focus for the

scientific community. Recently, many researchers have spent important efforts on

the study of protein-protein interactions (PPIs) in pathogen Trypanosome species

concluding that the low sequence identities between some parasite proteins and

their human host render these PPIs as highly promising drug targets. To the best

of our knowledge, there are no general models to predict Unique PPIs in

Trypanosome (TPPIs). On the other hand, the 3D structure of an increasing number

of Trypanosome proteins is reported in databases. In this regard, the

introduction of a new model to predict TPPIs from the 3D structure of proteins

involved in PPI is very important. For this purpose, we introduced new

protein-protein complex invariants based on the Markov average electrostatic

potential xi(k)(R(i)) for amino acids located in different regions (R(i)) of i-th

protein and placed at a distance k one from each other. We calculated more than

30 different types of parameters for 7866 pairs of proteins (1023 TPPIs and 6823

non-TPPIs) from more than 20 organisms, including parasites and human or cattle

hosts. We found a very simple linear model that predicts above 90% of TPPIs and

non-TPPIs both in training and independent test subsets using only two

parameters. The parameters were (d)xi(k)(s) = |xi(k)(s(1)) - xi(k)(s(2))|, the

absolute difference between the xi(k)(s(i)) values on the surface of the two

proteins of the pairs. We also tested nonlinear ANN models for comparison

purposes but the linear model gives the best results. We implemented this

predictor in the web server named TrypanoPPI freely available to public at

http://miaja.tic.udc.es/Bio-AIMS/TrypanoPPI.php. This is the first model that

predicts how unique a protein-protein complex in Trypanosome proteome is with

respect to other parasites and hosts, opening new opportunities for

antitrypanosome drug target discovery.

DOI: 10.1021/pr900827b

PMID: 19947655 [Indexed for MEDLINE]

2057. Bioinformatics. 2010 Feb 1;26(3):408-10. doi: 10.1093/bioinformatics/btp689. Epub

2009 Dec 18.

webMGR: an online tool for the multiple genome rearrangement problem.

Lin CH(1), Zhao H, Lowcay SH, Shahab A, Bourque G.

Author information:

(1)Genome Institute of Singapore, 60 Biopolis Street, #02-01, Genome, Singapore

138672.

SUMMARY: The algorithm MGR enables the reconstruction of rearrangement

phylogenies based on gene or synteny block order in multiple genomes. Although

MGR has been successfully applied to study the evolution of different sets of

species, its utilization has been hampered by the prohibitive running time for

some applications. In the current work, we have designed new heuristics that

significantly speed up the tool without compromising its accuracy. Moreover, we

have developed a web server (webMGR) that includes elaborate web output to

facilitate navigation through the results.

AVAILABILITY: webMGR can be accessed via http://www.gis.a-star.edu.sg/~bourque.

The source code of the improved standalone version of MGR is also freely

available from the web site.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp689

PMID: 20022974 [Indexed for MEDLINE]

2058. Bioinformatics. 2010 Feb 1;26(3):326-32. doi: 10.1093/bioinformatics/btp691. Epub

2009 Dec 17.

FoldAmyloid: a method of prediction of amyloidogenic regions from protein

sequence.

Garbuzynskiy SO(1), Lobanov MY, Galzitskaya OV.

Author information:

(1)Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow

Region 142290, Russia.

MOTIVATION: Amyloidogenic regions in polypeptide chains are very important

because such regions are responsible for amyloid formation and aggregation. It is

useful to be able to predict positions of amyloidogenic regions in protein

chains.

RESULTS: Two characteristics (expected probability of hydrogen bonds formation

and expected packing density of residues) have been introduced by us to detect

amyloidogenic regions in a protein sequence. We demonstrate that regions with

high expected probability of the formation of backbone-backbone hydrogen bonds as

well as regions with high expected packing density are mostly responsible for the

formation of amyloid fibrils. Our method (FoldAmyloid) has been tested on a

dataset of 407 peptides (144 amyloidogenic and 263 non-amyloidogenic peptides)

and has shown good performance in predicting a peptide status: amyloidogenic or

non-amyloidogenic. The prediction based on the expected packing density

classified correctly 75% of amyloidogenic peptides and 74% of non-amyloidogenic

ones. Two variants (averaging by donors and by acceptors) of prediction based on

the probability of formation of backbone-backbone hydrogen bonds gave a

comparable efficiency. With a hybrid-scale constructed by merging the above three

scales, our method is correct for 80% of amyloidogenic peptides and for 72% of

non-amyloidogenic ones. Prediction of amyloidogenic regions in proteins where

positions of amyloidogenic regions are known from experimental data has also been

done. In the proteins, our method correctly finds 10 out of 11 amyloidogenic

regions.

AVAILABILITY: The FoldAmyloid server is available at

http://antares.protres.ru/fold-amyloid/.

DOI: 10.1093/bioinformatics/btp691

PMID: 20019059 [Indexed for MEDLINE]

2059. Nan Fang Yi Ke Da Xue Xue Bao. 2010 Feb;30(2):219-23.

[Design and construction of the platform for comparative genomics].

[Article in Chinese]

Liu N(1), Guo YB, Sun XH, Ma L, Deng QK.

Author information:

(1)Institute of Bioinformatics, School of Basic Medical Science, College of

Biotechnology, Southern Medical University, Guangzhou 510515, China.

liuna520@fimmu.com

OBJECTIVE: To design a versatile genome comparison and visualization platform

based on browser/server mode supported by a local server.

METHODS: The server of the platform was Apache HTTP server. Perl was used to

integrate such genome alignment package and algorithms as MUMmer, LAGAN, and

Mauve for different comparison purposes, and the users could submit data and

parameters to the platform via the web page. The results of analysis were also

returned via the web page.

RESULTS: The platform could handle multiple data input formats, compare complete

and draft genome sequence, alignment pair-wise or multi genome of more divergent

species, identify regions of high similarity, locate local nucleotide mutations

and large-scale recombination, and display the results by visualization

interface. Analysis of the homology of 10 new strains of influenza A virus

indicated that PB1 gene might evolve from human H3N2 viruses, PB2 and PA genes

from avian H3N2 viruses, and HA and NS genes from swine H1N1 viruses. Alignment

of Mycobacterium tuberculosis (H37Rv, CDC1551) and Mycobacterium bovis

(AF2122/97) genomes revealed that sequence insertion/deletion and duplication

were the major source of genomic differences.

CONCLUSION: The platform integrate comprehensive resources with a user-friendly

interface and intuitive result visualization to facilitate conventional study of

comparative genomics.

PMID: 20159684 [Indexed for MEDLINE]

2060. Proteins. 2010 Feb 1;78(2):309-24. doi: 10.1002/prot.22544.

RigidFinder: a fast and sensitive method to detect rigid blocks in large

macromolecular complexes.

Abyzov A(1), Bjornson R, Felipe M, Gerstein M.

Author information:

(1)Department of Molecular Biophysics and Biochemistry, Yale University, New

Haven, Connecticut 06520, USA.

Advances in structure determination have made possible the analysis of large

macromolecular complexes (some with nearly 10,000 residues, such as GroEL). The

large-scale conformational changes associated with these complexes require new

approaches. Historically, a crucial component of motion analysis has been the

identification of moving rigid blocks from the comparison of different

conformations. However, existing tools do not allow consistent block

identification in very large structures. Here, we describe a novel method,

RigidFinder, for such identification of rigid blocks from different

conformations-across many scales, from large complexes to small loops.

RigidFinder defines rigidity in terms of blocks, where inter-residue distances

are conserved across conformations. Distance conservation, unlike the averaged

values (e.g., RMSD) used by many other methods, allows for sensitive

identification of motions. A further distinguishing feature of our method, is

that, it is capable of finding blocks made from nonconsecutive fragments of

multiple polypeptide chains. In our implementation, we utilize an efficient

quasi-dynamic programming search algorithm that allows for real-time application

to very large structures. RigidFinder can be used at a dedicated web server

(http://rigidfinder.molmovdb.org). The server also provides links to examples at

various scales such as loop closure, domain motions, partial refolding, and

subunit shifts. Moreover, here we describe the detailed application of

RigidFinder to four large structures: Pyruvate Phosphate Dikinase, T7 RNA

polymerase, RNA polymerase II, and GroEL. The results of the method are in

excellent agreement with the expert-described rigid blocks.

(c) 2009 Wiley-Liss, Inc.

DOI: 10.1002/prot.22544

PMID: 19705487 [Indexed for MEDLINE]

2061. BMC Bioinformatics. 2010 Jan 28;11:62. doi: 10.1186/1471-2105-11-62.

A novel scoring function for discriminating hyperthermophilic and mesophilic

proteins with application to predicting relative thermostability of protein

mutants.

Li Y(1), Middaugh CR, Fang J.

Author information:

(1)Applied Bioinformatics Laboratory, the University of Kansas, Lawrence, KS

66047, USA.

BACKGROUND: The ability to design thermostable proteins is theoretically

important and practically useful. Robust and accurate algorithms, however, remain

elusive. One critical problem is the lack of reliable methods to estimate the

relative thermostability of possible mutants.

RESULTS: We report a novel scoring function for discriminating hyperthermophilic

and mesophilic proteins with application to predicting the relative

thermostability of protein mutants. The scoring function was developed based on

an elaborate analysis of a set of features calculated or predicted from 540 pairs

of hyperthermophilic and mesophilic protein ortholog sequences. It was

constructed by a linear combination of ten important features identified by a

feature ranking procedure based on the random forest classification algorithm.

The weights of these features in the scoring function were fitted by a

hill-climbing algorithm. This scoring function has shown an excellent ability to

discriminate hyperthermophilic from mesophilic sequences. The prediction

accuracies reached 98.9% and 97.3% in discriminating orthologous pairs in

training and the holdout testing datasets, respectively. Moreover, the scoring

function can distinguish non-homologous sequences with an accuracy of 88.4%.

Additional blind tests using two datasets of experimentally investigated

mutations demonstrated that the scoring function can be used to predict the

relative thermostability of proteins and their mutants at very high accuracies

(92.9% and 94.4%). We also developed an amino acid substitution preference matrix

between mesophilic and hyperthermophilic proteins, which may be useful in

designing more thermostable proteins.

CONCLUSIONS: We have presented a novel scoring function which can distinguish not

only HP/MP ortholog pairs, but also non-homologous pairs at high accuracies. Most

importantly, it can be used to accurately predict the relative stability of

proteins and their mutants, as demonstrated in two blind tests. In addition, the

residue substitution preference matrix assembled in this study may reflect the

thermal adaptation induced substitution biases. A web server implementing the

scoring function and the dataset used in this study are freely available at

http://www.abl.ku.edu/thermorank/.

DOI: 10.1186/1471-2105-11-62

PMCID: PMC3098108

PMID: 20109199 [Indexed for MEDLINE]

2062. Stand Genomic Sci. 2010 Jan 28;2(1):142-8. doi: 10.4056/sigs.541628.

Standard operating procedure for calculating genome-to-genome distances based on

high-scoring segment pairs.

Auch AF, Klenk HP, Göker M.

DNA-DNA hybridization (DDH) is a widely applied wet-lab technique to obtain an

estimate of the overall similarity between the genomes of two organisms. To base

the species concept for prokaryotes ultimately on DDH was chosen by

microbiologists as a pragmatic approach for deciding about the recognition of

novel species, but also allowed a relatively high degree of standardization

compared to other areas of taxonomy. However, DDH is tedious and error-prone and

first and foremost cannot be used to incrementally establish a comparative

database. Recent studies have shown that in-silico methods for the comparison of

genome sequences can be used to replace DDH. Considering the ongoing rapid

technological progress of sequencing methods, genome-based prokaryote taxonomy is

coming into reach. However, calculating distances between genomes is dependent on

multiple choices for software and program settings. We here provide an overview

over the modifications that can be applied to distance methods based in

high-scoring segment pairs (HSPs) or maximally unique matches (MUMs) and that

need to be documented. General recommendations on determining HSPs using BLAST or

other algorithms are also provided. As a reference implementation, we introduce

the GGDC web server (http://ggdc.gbdp.org).

DOI: 10.4056/sigs.541628

PMCID: PMC3035261

PMID: 21304686

2063. BMC Bioinformatics. 2010 Jan 20;11:41. doi: 10.1186/1471-2105-11-41.

Prediction of the binding affinities of peptides to class II MHC using a

regularized thermodynamic model.

Bordner AJ(1), Mittelmann HD.

Author information:

(1)Mayo Clinic, Scottsdale, AZ 85259, USA. bordner.andrew@mayo.edu

BACKGROUND: The binding of peptide fragments of extracellular peptides to class

II MHC is a crucial event in the adaptive immune response. Each MHC allotype

generally binds a distinct subset of peptides and the enormous number of possible

peptide epitopes prevents their complete experimental characterization.

Computational methods can utilize the limited experimental data to predict the

binding affinities of peptides to class II MHC.

RESULTS: We have developed the Regularized Thermodynamic Average, or RTA, method

for predicting the affinities of peptides binding to class II MHC. RTA accounts

for all possible peptide binding conformations using a thermodynamic average and

includes a parameter constraint for regularization to improve accuracy on novel

data. RTA was shown to achieve higher accuracy, as measured by AUC, than

SMM-align on the same data for all 17 MHC allotypes examined. RTA also gave the

highest accuracy on all but three allotypes when compared with results from 9

different prediction methods applied to the same data. In addition, the method

correctly predicted the peptide binding register of 17 out of 18 peptide-MHC

complexes. Finally, we found that suboptimal peptide binding registers, which are

often ignored in other prediction methods, made significant contributions of at

least 50% of the total binding energy for approximately 20% of the peptides.

CONCLUSIONS: The RTA method accurately predicts peptide binding affinities to

class II MHC and accounts for multiple peptide binding registers while reducing

overfitting through regularization. The method has potential applications in

vaccine design and in understanding autoimmune disorders. A web server

implementing the RTA prediction method is available at

http://bordnerlab.org/RTA/.

DOI: 10.1186/1471-2105-11-41

PMCID: PMC2828437

PMID: 20089173 [Indexed for MEDLINE]

2064. BMC Bioinformatics. 2010 Jan 18;11 Suppl 1:S9. doi: 10.1186/1471-2105-11-S1-S9.

Prediction of protein structural classes for low-homology sequences based on

predicted secondary structure.

Yang JY(1), Peng ZL, Chen X.

Author information:

(1)Division of Mathematical Sciences, School of Physical and Mathematical

Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore.

yang0241@ntu.edu.sg

BACKGROUND: Prediction of protein structural classes (alpha, beta, alpha + beta

and alpha/beta) from amino acid sequences is of great importance, as it is

beneficial to study protein function, regulation and interactions. Many methods

have been developed for high-homology protein sequences, and the prediction

accuracies can achieve up to 90%. However, for low-homology sequences whose

average pairwise sequence identity lies between 20% and 40%, they perform

relatively poorly, yielding the prediction accuracy often below 60%.

RESULTS: We propose a new method to predict protein structural classes on the

basis of features extracted from the predicted secondary structures of proteins

rather than directly from their amino acid sequences. It first uses PSIPRED to

predict the secondary structure for each protein sequence. Then, the chaos game

representation is employed to represent the predicted secondary structure as two

time series, from which we generate a comprehensive set of 24 features using

recurrence quantification analysis, K-string based information entropy and

segment-based analysis. The resulting feature vectors are finally fed into a

simple yet powerful Fisher's discriminant algorithm for the prediction of protein

structural classes. We tested the proposed method on three benchmark datasets in

low homology and achieved the overall prediction accuracies of 82.9%, 83.1% and

81.3%, respectively. Comparisons with ten existing methods showed that our method

consistently performs better for all the tested datasets and the overall accuracy

improvements range from 2.3% to 27.5%. A web server that implements the proposed

method is freely available at http://www1.spms.ntu.edu.sg/~chenxin/RKS\_PPSC/.

CONCLUSION: The high prediction accuracy achieved by our proposed method is

attributed to the design of a comprehensive feature set on the predicted

secondary structure sequences, which is capable of characterizing the sequence

order information, local interactions of the secondary structural elements, and

spacial arrangements of alpha helices and beta strands. Thus, it is a valuable

method to predict protein structural classes particularly for low-homology amino

acid sequences.

DOI: 10.1186/1471-2105-11-S1-S9

PMCID: PMC3009544

PMID: 20122246 [Indexed for MEDLINE]

2065. BMC Bioinformatics. 2010 Jan 18;11 Suppl 1:S24. doi: 10.1186/1471-2105-11-S1-S24.

HORI: a web server to compute Higher Order Residue Interactions in protein

structures.

Sundaramurthy P(1), Shameer K, Sreenivasan R, Gakkhar S, Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences (TIFR), GKVK Campus, Bellary Road,

Bangalore, 560065, India. pksmurthy@ncbs.res.in

BACKGROUND: Folding of a protein into its three dimensional structure is

influenced by both local and global interactions within a protein. Higher order

residue interactions, like pairwise, triplet and quadruplet ones, play a vital

role in attaining the stable conformation of the protein structure. It is

generally agreed that higher order interactions make significant contribution to

the potential energy landscape of folded proteins and therefore it is important

to identify them to estimate their contributions to overall stability of a

protein structure.

RESULTS: We developed HORI [Higher order residue interactions in proteins], a web

server for the calculation of global and local higher order interactions in

protein structures. The basic algorithm of HORI is designed based on the

classical concept of four-body nearest-neighbour propensities of amino-acid

residues. It has been proved that higher order residue interactions up to the

level of quadruple interactions plays a major role in the three-dimensional

structure of proteins and is an important feature that can be used in protein

structure analysis.

CONCLUSION: HORI server will be a useful resource for the structural

bioinformatics community to perform analysis on protein structures based on

higher order residue interactions. HORI server is a highly interactive web server

designed in three modules that enables the user to analyse higher order residue

interactions in protein structures. HORI server is available from the URL:

http://caps.ncbs.res.in/hori.

DOI: 10.1186/1471-2105-11-S1-S24

PMCID: PMC3009495

PMID: 20122196 [Indexed for MEDLINE]

2066. BMC Plant Biol. 2010 Jan 18;10:14. doi: 10.1186/1471-2229-10-14.

SoyDB: a knowledge database of soybean transcription factors.

Wang Z(1), Libault M, Joshi T, Valliyodan B, Nguyen HT, Xu D, Stacey G, Cheng J.

Author information:

(1)Computer Science Department, University of Missouri, Columbia, MO 65211, USA.

BACKGROUND: Transcription factors play the crucial rule of regulating gene

expression and influence almost all biological processes. Systematically

identifying and annotating transcription factors can greatly aid further

understanding their functions and mechanisms. In this article, we present SoyDB,

a user friendly database containing comprehensive knowledge of soybean

transcription factors.

DESCRIPTION: The soybean genome was recently sequenced by the Department of

Energy-Joint Genome Institute (DOE-JGI) and is publicly available. Mining of this

sequence identified 5,671 soybean genes as putative transcription factors. These

genes were comprehensively annotated as an aid to the soybean research community.

We developed SoyDB - a knowledge database for all the transcription factors in

the soybean genome. The database contains protein sequences, predicted tertiary

structures, putative DNA binding sites, domains, homologous templates in the

Protein Data Bank (PDB), protein family classifications, multiple sequence

alignments, consensus protein sequence motifs, web logo of each family, and web

links to the soybean transcription factor database PlantTFDB, known EST

sequences, and other general protein databases including Swiss-Prot, Gene

Ontology, KEGG, EMBL, TAIR, InterPro, SMART, PROSITE, NCBI, and Pfam. The

database can be accessed via an interactive and convenient web server, which

supports full-text search, PSI-BLAST sequence search, database browsing by

protein family, and automatic classification of a new protein sequence into one

of 64 annotated transcription factor families by hidden Markov models.

CONCLUSIONS: A comprehensive soybean transcription factor database was

constructed and made publicly accessible at http://casp.rnet.missouri.edu/soydb/.

DOI: 10.1186/1471-2229-10-14

PMCID: PMC2826334

PMID: 20082720 [Indexed for MEDLINE]

2067. Bioinformatics. 2010 Jan 15;26(2):273-4. doi: 10.1093/bioinformatics/btp646. Epub

2009 Nov 25.

CMap3D: a 3D visualization tool for comparative genetic maps.

Duran C(1), Boskovic Z, Imelfort M, Batley J, Hamilton NA, Edwards D.

Author information:

(1)Australian Centre for Plant Functional Genomics, University of Queensland,

Brisbane, QLD 4072, Australia.

Genetic linkage mapping enables the study of genome organization and the

association of heritable traits with regions of sequenced genomes. Comparative

genetic mapping is particularly powerful as it allows translation of information

between related genomes and gives an insight into genome evolution. A common tool

for the storage, comparison and visualization of genetic maps is CMap. However,

current visualization in CMap is limited to the comparison of adjacent aligned

maps. To overcome this limitation, we have developed CMap3D, a tool to compare

multiple genetic maps in three-dimensional space. CMap3D is based on a

client/server model ensuring operability with current CMap data repositories.

This tool can be applied to any species where genetic map information is

available and enables rapid, direct comparison between multiple aligned

maps.AVAILABILITY AND IMPLEMENTATION: The software is a stand-alone application

written in Processing and Java. Binaries are available for Windows, OSX and

Linux, and require Sun Microsystems Java Runtime Environment 1.6 or later. The

software is freely available for non-commercial use from

http://flora.acpfg.com.au/.

DOI: 10.1093/bioinformatics/btp646

PMID: 19942584 [Indexed for MEDLINE]

2068. J Comput Chem. 2010 Jan 15;31(1):217-23. doi: 10.1002/jcc.21281.

Prediction of membrane spanning segments and topology in beta-barrel membrane

proteins at better accuracy.

Ou YY(1), Chen SA, Gromiha MM.

Author information:

(1)Department of Computer Science and Engineering, Yuan Ze University, Chung-Li,

Taiwan.

Prediction of membrane spanning segments in beta-barrel outer membrane proteins

(OMP) and their topology is an important problem in structural and functional

genomics. In this work, we propose a method based on radial basis networks for

predicting the number of beta-strands in OMPs and identifying their membrane

spanning segments. Our method showed a leave-one-out cross validation accuracy of

96% in a set of 28 OMPs, which have the range of 8-22 beta-strand segments. The

beta-strand segments in OMPs and the residues in membrane spanning segments are

correctly predicted with the accuracy of 96% and 87%, respectively. We have

developed a web server, TMBETAPRED-RBF for predicting the transmembrane

beta-strands from amino acid sequence and it is available at

http://rbf.bioinfo.tw/~sachen/tmrbf.html. We suggest that our method could be an

effective tool for predicting the membrane spanning regions and topology of

beta-barrel membrane proteins.

Copyright 2009 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.21281

PMID: 19421989 [Indexed for MEDLINE]

2069. BMC Evol Biol. 2010 Jan 12;10:8. doi: 10.1186/1471-2148-10-8.

BLAST-EXPLORER helps you building datasets for phylogenetic analysis.

Dereeper A(1), Audic S, Claverie JM, Blanc G.

Author information:

(1)Information Génomique & Structurale-CNRS-UPR2589, Institut de Microbiologie de

la Méditerranée-IFR 88, Marseille, France.

BACKGROUND: The right sampling of homologous sequences for phylogenetic or

molecular evolution analyses is a crucial step, the quality of which can have a

significant impact on the final interpretation of the study. There is no single

way for constructing datasets suitable for phylogenetic analysis, because this

task intimately depends on the scientific question we want to address, Moreover,

database mining softwares such as BLAST which are routinely used for searching

homologous sequences are not specifically optimized for this task.

RESULTS: To fill this gap, we designed BLAST-Explorer, an original and friendly

web-based application that combines a BLAST search with a suite of tools that

allows interactive, phylogenetic-oriented exploration of the BLAST results and

flexible selection of homologous sequences among the BLAST hits. Once the

selection of the BLAST hits is done using BLAST-Explorer, the corresponding

sequence can be imported locally for external analysis or passed to the

phylogenetic tree reconstruction pipelines available on the Phylogeny.fr

platform.

CONCLUSIONS: BLAST-Explorer provides a simple, intuitive and interactive

graphical representation of the BLAST results and allows selection and retrieving

of the BLAST hit sequences based a wide range of criterions. Although

BLAST-Explorer primarily aims at helping the construction of sequence datasets

for further phylogenetic study, it can also be used as a standard BLAST server

with enriched output. BLAST-Explorer is available at http://www.phylogeny.fr.

DOI: 10.1186/1471-2148-10-8

PMCID: PMC2821324

PMID: 20067610 [Indexed for MEDLINE]

2070. Bioinformation. 2010 Jan 7;4(7):271-5.

FAIR: A server for internal sequence repeats.

Senthilkumar R(1), Sabarinathan R, Hameed BS, Banerjee N, Chidambarathanu N,

Karthik R, Sekar K.

Author information:

(1)Bioinformatics Centre, Centre of Excellence in Structural Biology and

Bio-computing, Indian Institute of Science, Bangalore 560 012, India.

An Internet computing server has been developed to identify all the occurrences

of the internal sequence repeats in a protein and DNA sequences. Further, an

option is provided for the users to check the occurrence(s) of the resultant

sequence repeats in the other sequence and structure (Protein Data Bank)

databases. The databases deployed in the proposed computing engine are up-to-date

and thus the users will get the latest information available in the respective

databases. The server is freely accessible over the World Wide Web

(WWW).AVAILABILITY: http://bioserver1.physics.iisc.ernet.in/fair/

PMCID: PMC2957760

PMID: 20978598

2071. Adv Exp Med Biol. 2010;680:361-9. doi: 10.1007/978-1-4419-5913-3\_41.

An overview of the BioExtract Server: a distributed, Web-based system for genomic

analysis.

Lushbough CM(1), Brendel VP.

Author information:

(1)Department of Computer Science, University of South Dakota, Vermillion, SD,

USA. carol.lushbough@usd.edu

Genome research is becoming increasingly dependent on access to multiple,

distributed data sources, and bioinformatic tools. The importance of integration

across distributed databases and Web services will continue to grow as the number

of requisite resources expands. Use of bioinformatic workflows has seen

considerable growth in recent years as scientific research becomes increasingly

dependent on the analysis of large sets of data and the use of distributed

resources. The BioExtract Server (http://bioextract.org) is a Web-based system

designed to aid researchers in the analysis of distributed genomic data by

providing a platform to facilitate the creation of bioinformatic workflows.

Scientific workflows are created within the system by recording the analytic

tasks preformed by researchers. These steps may include querying multiple data

sources, saving query results as searchable data extracts, and executing local

and Web-accessible analytic tools. The series of recorded tasks can be saved as a

computational workflow simply by providing a name and description.

DOI: 10.1007/978-1-4419-5913-3\_41

PMID: 20865520 [Indexed for MEDLINE]

2072. Bioinformatics. 2010 Jan 1;26(1):130-1. doi: 10.1093/bioinformatics/btp600. Epub

2009 Oct 22.

Ancestors 1.0: a web server for ancestral sequence reconstruction.

Diallo AB(1), Makarenkov V, Blanchette M.

Author information:

(1)Department of Computer Science, Université du Québec à Montréal, PO. Box 8888

Downtown Station, Montreal, QC H3C3P8, Canada. diallo.abdoulaye@uqam.ca

SUMMARY: The computational inference of ancestral genomes consists of five

difficult steps: identifying syntenic regions, inferring ancestral arrangement of

syntenic regions, aligning multiple sequences, reconstructing the insertion and

deletion history and finally inferring substitutions. Each of these steps have

received lot of attention in the past years. However, there currently exists no

framework that integrates all of the different steps in an easy workflow. Here,

we introduce Ancestors 1.0, a web server allowing one to easily and quickly

perform the last three steps of the ancestral genome reconstruction procedure. It

implements several alignment algorithms, an indel maximum likelihood solver and a

context-dependent maximum likelihood substitution inference algorithm. The

results presented by the server include the posterior probabilities for the last

two steps of the ancestral genome reconstruction and the expected error rate of

each ancestral base prediction.

AVAILABILITY: The Ancestors 1.0 is available at

http://ancestors.bioinfo.uqam.ca/ancestorWeb/.

DOI: 10.1093/bioinformatics/btp600

PMID: 19850756 [Indexed for MEDLINE]

2073. Conf Proc IEEE Eng Med Biol Soc. 2010;2010:1494-7. doi:

10.1109/IEMBS.2010.5626843.

WebPK, a web-based tool for custom pharmacokinetic simulation.

Srimani J(1), Moffitt RA, Wang MD.

Author information:

(1)department of Electrical and Computer Engineering, Georgia Institute of

Technology, Atlanta, GA 30332, USA. jaydeep.srimani@gatech.edu

Drug bioavailability is a major failing point of new pharmaceuticals i.e. drugs

fail to reach their target or fail to stay there long enough for therapeutic

effect. Compounding this issue, significant variability exists between patients

and how they metabolize and distribute a drug. We present WebPK, a web-based tool

for simulation of custom pharmacokinetic models. Model parameters can be entered

manually or uploaded as a file. Simulation computations are performed on the

server side, and thus require minimal client resources, which makes WebPK

suitable for mobile devices. Time series biodistribution data are returned to the

user in graphical and numerical form for quick interpretation or archiving.

Results generated from WebPK are consistent with previously published

pharmacokinetic models. This work is expected to provide physicians with access

to easy simulation of patient pharmacokinetic profiles, which will allow for the

prescription of more efficient and personalized drug regimens. URL:

http://webpk.bme.gatech.edu.

DOI: 10.1109/IEMBS.2010.5626843

PMCID: PMC5003049

PMID: 21096365 [Indexed for MEDLINE]

2074. Curr Pharm Des. 2010;16(24):2710-23.

Review of QSAR models for enzyme classes of drug targets: Theoretical background

and applications in parasites, hosts, and other organisms.

Concu R(1), Podda G, Ubeira FM, González-Díaz H.

Author information:

(1)Department of Microbiology and Parasitology, University of Santiago de

Compostela, Spain.

The number of protein 3D structures without function annotation in Protein Data

Bank (PDB) has been steadily increased. Many of these proteins are relevant for

Pharmaceutical Design because they may be enzymes of different classes that could

become drug targets. This fact has led in turn to an increment of demand for

theoretical models to give a quick characterization of these proteins. In this

work, we present a review and discussion of Alignment-Free Methods (AFMs) for

fast prediction of the Enzyme Classification (EC) number from structural

patterns. We referred to both methods based on linear techniques such as Linear

Discriminant Analysis (LDA) and/or non-linear models like Artificial Neural

Networks (ANN) or Support Vector Machine (SVM) in order to compare linear vs.

non-linear classifiers. We also detected which of these models have been

implemented as Web Servers free to the public and compiled a list of some of

these web sites. For instance, we reviewed the servers implemented at portal

Bio-AIMS (http://miaja.tic.udc.es/Bio-AIMS/EnzClassPred.php) and the server

EzyPred (http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/).

PMID: 20642430 [Indexed for MEDLINE]

2075. Database (Oxford). 2010;2010:bap025. doi: 10.1093/database/bap025. Epub 2010 Jan

8.

CyanoClust: comparative genome resources of cyanobacteria and plastids.

Sasaki NV(1), Sato N.

Author information:

(1)Department of Life Sciences, Graduate School of Arts and Sciences, University

of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo, 153-8902, Japan.

naobumi@bio.c.u-tokyo.ac.jp

Cyanobacteria, which perform oxygen-evolving photosynthesis as do chloroplasts of

plants and algae, are one of the best-studied prokaryotic phyla and one from

which many representative genomes have been sequenced. Lack of a suitable

comparative genomic database has been a problem in cyanobacterial genomics

because many proteins involved in physiological functions such as photosynthesis

and nitrogen fixation are not catalogued in commonly used databases, such as

Clusters of Orthologous Proteins (COG). CyanoClust is a database of homolog

groups in cyanobacteria and plastids that are produced by the program Gclust. We

have developed a web-server system for the protein homology database featuring

cyanobacteria and plastids. Database URL: http://cyanoclust.c.u-tokyo.ac.jp/.

DOI: 10.1093/database/bap025

PMCID: PMC2860898

PMID: 20428314 [Indexed for MEDLINE]

2076. IEEE/ACM Trans Comput Biol Bioinform. 2010 Jan-Mar;7(1):12-24. doi:

10.1109/TCBB.2008.98.

BioExtract server--an integrated workflow-enabling system to access and analyze

heterogeneous, distributed biomolecular data.

Lushbough C(1), Bergman MK, Lawrence CJ, Jennewein D, Brendel V.

Author information:

(1)Department of Computer Science, University of South Dakota, 414 East Clark

St., Vermillion, SD 57069, USA. Carol.Lushbough@usd.edu

Many in silico investigations in bioinformatics require access to multiple,

distributed data sources and analytic tools. The requisite data sources may

include large public data repositories, community databases, and project

databases for use in domain-specific research. Different data sources frequently

utilize distinct query languages and return results in unique formats, and

therefore researchers must either rely upon a small number of primary data

sources or become familiar with multiple query languages and formats. Similarly,

the associated analytic tools often require specific input formats and produce

unique outputs which make it difficult to utilize the output from one tool as

input to another. The BioExtract Server (http://bioextract.org) is a Web-based

data integration application designed to consolidate, analyze, and serve data

from heterogeneous biomolecular databases in the form of a mash-up. The basic

operations of the BioExtract Server allow researchers, via their Web browsers, to

specify data sources, flexibly query data sources, apply analytic tools, download

result sets, and store query results for later reuse. As a researcher works with

the system, their "steps" are saved in the background. At any time, these steps

can be preserved long-term as a workflow simply by providing a workflow name and

description.

DOI: 10.1109/TCBB.2008.98

PMID: 20150665 [Indexed for MEDLINE]

2077. In Silico Biol. 2010;10(3):185-91. doi: 10.3233/ISB-2010-0426.

POODLE-I: disordered region prediction by integrating POODLE series and

structural information predictors based on a workflow approach.

Hirose S(1), Shimizu K, Noguchi T.

Author information:

(1)Computational Biology Research Center (CBRC), National Institute of Advanced

Industrial Science and Technology (AIST), Tokyo, Japan. hirose-shuichi@aist.go.jp

Under physiological conditions, many proteins that include a region lacking

well-defined three-dimensional structures have been identified, especially in

eukaryotes. These regions often play an important biological cellular role,

although they cannot form a stable structure. Therefore, they are biologically

remarkable phenomena. From an industrial perspective, they can provide useful

information for determining three-dimensional structures or designing drugs. For

these reasons, disordered regions have attracted a great deal of attention in

recent years. Their accurate prediction is therefore anticipated to provide

annotations that are useful for wide range of applications. POODLE-I (where "I"

stands for integration) is a web-based disordered region prediction system.

POODLE-I integrates prediction results obtained from three kinds of disordered

region predictors (POODLEs) developed from the viewpoint that the characteristics

of disordered regions change according to their length. Furthermore, POODLE-I

combines that information with predicted structural information by application of

a workflow approach. When compared with server teams that showed best performance

in CASP8, POODLE-I ranked among the top and exhibited the highest performance in

predicting unfolded proteins. POODLE-I is an efficient tool for detecting

disordered regions in proteins solely from the amino acid sequence. The

application is freely available at http://mbs.cbrc.jp/poodle/poodle-i.html.

DOI: 10.3233/ISB-2010-0426

PMID: 22430291 [Indexed for MEDLINE]

2078. J Comput Biol. 2010 Jan;17(1):73-7. doi: 10.1089/cmb.2009.0006.

GBA manager: an online tool for querying low-complexity regions in proteins.

Bandyopadhyay N(1), Kahveci T.

Author information:

(1)Department of Computer and Information Science and Engineering, University of

Florida, Gainesville, Florida, USA.

Abstract We developed GBA Manager, an online software that facilitates the

Graph-Based Algorithm (GBA) we proposed in our earlier work. GBA identifies the

low-complexity regions (LCR) of protein sequences. GBA exploits a similarity

matrix, such as BLOSUM62, to compute the complexity of the subsequences of the

input protein sequence. It uses a graph-based algorithm to accurately compute the

regions that have low complexities. GBA Manager is a user friendly web-service

that enables online querying of protein sequences using GBA. In addition to

querying capabilities of the existing GBA algorithm, GBA Manager computes the

p-values of the LCR identified. The p-value gives an estimate of the possibility

that the region appears by chance. GBA Manager presents the output in three

different understandable formats. GBA Manager is freely accessible at

http://bioinformatics.cise.ufl.edu/GBA/GBA.htm .

DOI: 10.1089/cmb.2009.0006

PMID: 20078398 [Indexed for MEDLINE]

2079. J Obstet Gynaecol. 2010;30(7):671-4. doi: 10.3109/01443615.2010.511323.

Intrapartum cardiotocograph interpretation by midwives and trainee obstetricians

using a modified definition of a fetal heart rate deceleration.

Gebeh A(1), Yulia A, Ayuk P.

Author information:

(1)Directorate of Women's Services, Royal Victoria Infirmary,

Newcastle-upon-Tyne, UK. alphagebeh@doctors.org.uk

The aim of this study was to assess cardiotocograph (CTG) interpretation by

midwives and trainee obstetricians using the standard and a modified definition

of fetal heart rate deceleration compared with consultant interpretation as the

Gold Standard. A randomised survey using the online tool at: www.surveymonkey.com

between 4 January and 24 April 2009, was conducted at a tertiary obstetric unit,

UK. A total of 104 (∼54%) health professionals responded, providing 1,118

responses with respect to the presence of decelerations on 13 anonymised CTGs.

Five obstetric consultants (62.5%) provided expert opinion. Midwives and trainee

obstetricians were more likely to concur with Consultant opinion when using the

modified definition of fetal heart rate deceleration compared with the standard

definition. Larger scale studies may be needed to further evaluate the usefulness

of the modified definition.

DOI: 10.3109/01443615.2010.511323

PMID: 20925607 [Indexed for MEDLINE]

2080. Methods Mol Biol. 2010;629:109-21. doi: 10.1007/978-1-60761-657-3\_8.

Using OligoWalk to identify efficient siRNA sequences.

Mathews DH(1).

Author information:

(1)Department of Biochemistry, University of Rochester Medical Center, Rochester,

NY, USA.

RNA interference (RNAi) has emerged as an important tool in science and in

medicine. Small-interfering RNAs (siRNAs) can be used to knockdown gene

expression of specific mRNAs. In practice, a number of factors influence whether

an siRNA sequence will elicit RNAi and knockdown target gene expression. One

factor that significantly influences the efficiency of an siRNA is the effect of

RNA secondary structure. Self-structure in either the siRNA sequence or the

target mRNA at the binding site may prevent gene silencing. This chapter provides

protocols for using the OligoWalk software package to design efficient siRNAs.

OligoWalk considers the effect of target and guide strand self-structures and

also local sequence features in siRNA design. OligoWalk can be run either locally

by compiling the software or through a convenient web interface. OligoWalk is

freely available at

http://rna.urmc.rochester.edu/cgi-bin/server\_exe/oligowalk/oligowalk\_form.cgi .

DOI: 10.1007/978-1-60761-657-3\_8

PMID: 20387146 [Indexed for MEDLINE]

2081. Nucleic Acids Res. 2010 Jan;38(Database issue):D620-5. doi: 10.1093/nar/gkp961.

Epub 2009 Nov 17.

ENCODE whole-genome data in the UCSC Genome Browser.

Rosenbloom KR(1), Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ,

Wang T, Hinrichs AS, Zweig AS, Fujita PA, Learned K, Rhead B, Smith KE, Kuhn RM,

Karolchik D, Haussler D, Kent WJ.

Author information:

(1)Center for Biomolecular Science and Engineering, School of Engineering,

University of California, Santa Cruz, CA 95064, USA. kate@soe.ucsc.edu

The Encyclopedia of DNA Elements (ENCODE) project is an international consortium

of investigators funded to analyze the human genome with the goal of producing a

comprehensive catalog of functional elements. The ENCODE Data Coordination Center

at The University of California, Santa Cruz (UCSC) is the primary repository for

experimental results generated by ENCODE investigators. These results are

captured in the UCSC Genome Bioinformatics database and download server for

visualization and data mining via the UCSC Genome Browser and companion tools

(Rhead et al. The UCSC Genome Browser Database: update 2010, in this issue). The

ENCODE web portal at UCSC (http://encodeproject.org or

http://genome.ucsc.edu/ENCODE) provides information about the ENCODE data and

convenient links for access.

DOI: 10.1093/nar/gkp961

PMCID: PMC2808953

PMID: 19920125 [Indexed for MEDLINE]

2082. Nucleic Acids Res. 2010 Jan;38(Database issue):D275-9. doi: 10.1093/nar/gkp966.

Epub 2009 Nov 17.

The ITS2 Database III--sequences and structures for phylogeny.

Koetschan C(1), Förster F, Keller A, Schleicher T, Ruderisch B, Schwarz R, Müller

T, Wolf M, Schultz J.

Author information:

(1)Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland

97074 Wuerzburg, Germany.

The internal transcribed spacer 2 (ITS2) is a widely used phylogenetic marker. In

the past, it has mainly been used for species level classifications. Nowadays, a

wider applicability becomes apparent. Here, the conserved structure of the RNA

molecule plays a vital role. We have developed the ITS2 Database

(http://its2.bioapps.biozentrum.uni-wuerzburg.de) which holds information about

sequence, structure and taxonomic classification of all ITS2 in GenBank. In the

new version, we use Hidden Markov models (HMMs) for the identification and

delineation of the ITS2 resulting in a major redesign of the annotation pipeline.

This allowed the identification of more than 160,000 correct full length and more

than 50,000 partial structures. In the web interface, these can now be searched

with a modified BLAST considering both sequence and structure, enabling rapid

taxon sampling. Novel sequences can be annotated using the HMM based approach and

modelled according to multiple template structures. Sequences can be searched for

known and newly identified motifs. Together, the database and the web server

build an exhaustive resource for ITS2 based phylogenetic analyses.

DOI: 10.1093/nar/gkp966

PMCID: PMC2808966

PMID: 19920122 [Indexed for MEDLINE]

2083. Nucleic Acids Res. 2010 Jan;38(Database issue):D167-80. doi: 10.1093/nar/gkp1016.

Epub 2009 Nov 17.

ELM: the status of the 2010 eukaryotic linear motif resource.

Gould CM(1), Diella F, Via A, Puntervoll P, Gemünd C, Chabanis-Davidson S,

Michael S, Sayadi A, Bryne JC, Chica C, Seiler M, Davey NE, Haslam N, Weatheritt

RJ, Budd A, Hughes T, Pas J, Rychlewski L, Travé G, Aasland R, Helmer-Citterich

M, Linding R, Gibson TJ.

Author information:

(1)Structural and Computational Biology Unit, European Molecular Biology

Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

Linear motifs are short segments of multidomain proteins that provide regulatory

functions independently of protein tertiary structure. Much of intracellular

signalling passes through protein modifications at linear motifs. Many thousands

of linear motif instances, most notably phosphorylation sites, have now been

reported. Although clearly very abundant, linear motifs are difficult to predict

de novo in protein sequences due to the difficulty of obtaining robust

statistical assessments. The ELM resource at http://elm.eu.org/ provides an

expanding knowledge base, currently covering 146 known motifs, with annotation

that includes >1300 experimentally reported instances. ELM is also an exploratory

tool for suggesting new candidates of known linear motifs in proteins of

interest. Information about protein domains, protein structure and native

disorder, cellular and taxonomic contexts is used to reduce or deprecate false

positive matches. Results are graphically displayed in a 'Bar Code' format, which

also displays known instances from homologous proteins through a novel 'Instance

Mapper' protocol based on PHI-BLAST. ELM server output provides links to the ELM

annotation as well as to a number of remote resources. Using the links,

researchers can explore the motifs, proteins, complex structures and associated

literature to evaluate whether candidate motifs might be worth experimental

investigation.

DOI: 10.1093/nar/gkp1016

PMCID: PMC2808914

PMID: 19920119 [Indexed for MEDLINE]

2084. Nucleic Acids Res. 2010 Jan;38(Database issue):D320-5. doi: 10.1093/nar/gkp1013.

Epub 2009 Nov 11.

Protein Geometry Database: a flexible engine to explore backbone conformations

and their relationships to covalent geometry.

Berkholz DS(1), Krenesky PB, Davidson JR, Karplus PA.

Author information:

(1)Department of Biochemistry and Biophysics, Oregon State University, 2011 ALS

and Open Source Lab, Oregon State University, B211 Kerr Admin, Corvallis OR

97331, USA.

The backbone bond lengths, bond angles, and planarity of a protein are influenced

by the backbone conformation (varphi,Psi), but no tool exists to explore these

relationships, leaving this area as a reservoir of untapped information about

protein structure and function. The Protein Geometry Database (PGD) enables

biologists to easily and flexibly query information about the conformation alone,

the backbone geometry alone, and the relationships between them. The capabilities

the PGD provides are valuable for assessing the uniqueness of observed

conformational or geometric features in protein structure as well as discovering

novel features and principles of protein structure. The PGD server is available

at http://pgd.science.oregonstate.edu/ and the data and code underlying it are

freely available to use and extend.

DOI: 10.1093/nar/gkp1013

PMCID: PMC2808862

PMID: 19906726 [Indexed for MEDLINE]

2085. Nucleic Acids Res. 2010 Jan;38(Database issue):D283-7. doi: 10.1093/nar/gkp963.

Epub 2009 Nov 11.

ComSin: database of protein structures in bound (complex) and unbound (single)

states in relation to their intrinsic disorder.

Lobanov MY(1), Shoemaker BA, Garbuzynskiy SO, Fong JH, Panchenko AR, Galzitskaya

OV.

Author information:

(1)Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow

Region, Russia.

Most of the proteins in a cell assemble into complexes to carry out their

function. In this work, we have created a new database (named ComSin) of protein

structures in bound (complex) and unbound (single) states to provide a researcher

with exhaustive information on structures of the same or homologous proteins in

bound and unbound states. From the complete Protein Data Bank (PDB), we selected

24 910 pairs of protein structures in bound and unbound states, and identified

regions of intrinsic disorder. For 2448 pairs, the proteins in bound and unbound

states are identical, while 7129 pairs have sequence identity 90% or larger. The

developed server enables one to search for proteins in bound and unbound states

with several options including sequence similarity between the corresponding

proteins in bound and unbound states, and validation of interaction interfaces of

protein complexes. Besides that, through our web server, one can obtain necessary

information for studying disorder-to-order and order-to-disorder transitions upon

complex formation, and analyze structural differences between proteins in bound

and unbound states. The database is available at

http://antares.protres.ru/comsin/.

DOI: 10.1093/nar/gkp963

PMCID: PMC2808974

PMID: 19906708 [Indexed for MEDLINE]

2086. Nucleic Acids Res. 2010 Jan;38(Database issue):D518-24. doi: 10.1093/nar/gkp842.

Epub 2009 Oct 20.

Inferred Biomolecular Interaction Server--a web server to analyze and predict

protein interacting partners and binding sites.

Shoemaker BA(1), Zhang D, Thangudu RR, Tyagi M, Fong JH, Marchler-Bauer A, Bryant

SH, Madej T, Panchenko AR.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, Bethesda, MD 20894, USA.

IBIS is the NCBI Inferred Biomolecular Interaction Server. This server organizes,

analyzes and predicts interaction partners and locations of binding sites in

proteins. IBIS provides annotations for different types of binding partners

(protein, chemical, nucleic acid and peptides), and facilitates the mapping of a

comprehensive biomolecular interaction network for a given protein query. IBIS

reports interactions observed in experimentally determined structural complexes

of a given protein, and at the same time IBIS infers binding sites/interacting

partners by inspecting protein complexes formed by homologous proteins. Similar

binding sites are clustered together based on their sequence and structure

conservation. To emphasize biologically relevant binding sites, several

algorithms are used for verification in terms of evolutionary conservation,

biological importance of binding partners, size and stability of interfaces, as

well as evidence from the published literature. IBIS is updated regularly and is

freely accessible via http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.html.

DOI: 10.1093/nar/gkp842

PMCID: PMC2808861

PMID: 19843613 [Indexed for MEDLINE]

2087. Nucleic Acids Res. 2010 Jan;38(Database issue):D570-6. doi: 10.1093/nar/gkp799.

Epub 2009 Sep 26.

EMMA--mouse mutant resources for the international scientific community.

Wilkinson P(1), Sengerova J, Matteoni R, Chen CK, Soulat G, Ureta-Vidal A,

Fessele S, Hagn M, Massimi M, Pickford K, Butler RH, Marschall S, Mallon AM,

Pickard A, Raspa M, Scavizzi F, Fray M, Larrigaldie V, Leyritz J, Birney E,

Tocchini-Valentini GP, Brown S, Herault Y, Montoliu L, de Angelis MH, Smedley D.

Author information:

(1)European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton,

Cambridge CB10 1SD, UK.

The laboratory mouse is the premier animal model for studying human disease and

thousands of mutants have been identified or produced, most recently through

gene-specific mutagenesis approaches. High throughput strategies by the

International Knockout Mouse Consortium (IKMC) are producing mutants for all

protein coding genes. Generating a knock-out line involves huge monetary and time

costs so capture of both the data describing each mutant alongside archiving of

the line for distribution to future researchers is critical. The European Mouse

Mutant Archive (EMMA) is a leading international network infrastructure for

archiving and worldwide provision of mouse mutant strains. It operates in

collaboration with the other members of the Federation of International Mouse

Resources (FIMRe), EMMA being the European component. Additionally EMMA is one of

four repositories involved in the IKMC, and therefore the current figure of 1700

archived lines will rise markedly. The EMMA database gathers and curates

extensive data on each line and presents it through a user-friendly website. A

BioMart interface allows advanced searching including integrated querying with

other resources e.g. Ensembl. Other resources are able to display EMMA data by

accessing our Distributed Annotation System server. EMMA database access is

publicly available at http://www.emmanet.org.

DOI: 10.1093/nar/gkp799

PMCID: PMC2808872

PMID: 19783817 [Indexed for MEDLINE]

2088. SAR QSAR Environ Res. 2010 Jan 1;21(1):37-55. doi: 10.1080/10629360903560637.

Analysis of hydrophobic interactions of antagonists with the beta2-adrenergic

receptor.

Novoseletsky VN(1), Pyrkov TV, Efremov RG.

Author information:

(1)MM Shemyakin & Yu A Ovchinnikov Institute of Bioorganic Chemistry, Russian

Academy of Sciences, Ul Miklukho-Maklaya, Moscow V-437, Russia. valeryns@nmr.ru

The adrenergic receptors mediate a wide variety of physiological responses,

including vasodilatation and vasoconstriction, heart rate modulation, and others.

Beta-adrenergic antagonists ('beta-blockers') thus constitute a widely used class

of drugs in cardiovascular medicine as well as in management of anxiety,

migraine, and glaucoma. The importance of the hydrophobic effect has been

evidenced for a wide range of beta-blocker properties. To better understand the

role of the hydrophobic effect in recognition of beta-blockers by their receptor,

we carried out a molecular docking study combined with an original approach to

estimate receptor-ligand hydrophobic interactions. The proposed method is based

on automatic detection of molecular fragments in ligands and the analysis of

their interactions with receptors separately. A series of beta-blockers, based on

phenylethanolamines and phenoxypropanolamines, were docked to the

beta2-adrenoceptor binding site in the crystal structure. Hydrophobic

complementarity between the ligand and the receptor was calculated using the

PLATINUM web-server (http://model.nmr.ru/platinum). Based on the analysis of the

hydrophobic match for molecular fragments of beta-blockers, we have developed a

new scoring function which efficiently predicts dissociation constant (pKd) with

strong correlations (r(2) approximately 0.8) with experimental data.

DOI: 10.1080/10629360903560637

PMID: 20373213 [Indexed for MEDLINE]

2089. Subst Abuse. 2010;4:21-33. Epub 2010 Sep 21.

Cellular phone-based image acquisition and quantitative ratiometric method for

detecting cocaine and benzoylecgonine for biological and forensic applications.

Cadle BA(1), Rasmus KC, Varela JA, Leverich LS, O'Neill CE, Bachtell RK, Cooper

DC.

Author information:

(1)Institute for Behavioral Genetics.

Here we describe the first report of using low-cost cellular or web-based digital

cameras to image and quantify standardized rapid immunoassay strips as a new

point-of-care diagnostic and forensics tool with health applications.

Quantitative ratiometric pixel density analysis (QRPDA) is an automated method

requiring end-users to utilize inexpensive (∼ $1 USD/each) immunotest strips, a

commonly available web or mobile phone camera or scanner, and internet or

cellular service. A model is described whereby a central computer server and

freely available IMAGEJ image analysis software records and analyzes the incoming

image data with time-stamp and geo-tag information and performs the QRPDA using

custom JAVA based macros (http://www.neurocloud.org). To demonstrate QRPDA we

developed a standardized method using rapid immunotest strips directed against

cocaine and its major metabolite, benzoylecgonine. Images from standardized

samples were acquired using several devices, including a mobile phone camera, web

cam, and scanner. We performed image analysis of three brands of commercially

available dye-conjugated anti-cocaine/benzoylecgonine (COC/BE) antibody test

strips in response to three different series of cocaine concentrations ranging

from 0.1 to 300 ng/ml and BE concentrations ranging from 0.003 to 0.1 ng/ml. This

data was then used to create standard curves to allow quantification of COC/BE in

biological samples. Across all devices, QRPDA quantification of COC and BE proved

to be a sensitive, economical, and faster alternative to more costly methods,

such as gas chromatography-mass spectrometry, tandem mass spectrometry, or high

pressure liquid chromatography. The limit of detection was determined to be

between 0.1 and 5 ng/ml. To simulate conditions in the field, QRPDA was found to

be robust under a variety of image acquisition and testing conditions that varied

temperature, lighting, resolution, magnification and concentrations of biological

fluid in a sample. To determine the effectiveness of the QRPDA method for

quantifying cocaine in biological samples, mice were injected with a

sub-locomotor activating dose of cocaine (5 mg/kg; i.p.) and were found to have

detectable levels of COC/BE in their urine (160.6 ng/ml) and blood plasma (8.1

ng/ml) after 15-30 minutes. By comparison rats self-administering cocaine in a 4

hour session obtained a final BE blood plasma level of 910 ng/ml with an average

of 62.5 infusions. It is concluded that automated QRPDA is a low-cost, rapid and

highly sensitive method for the detection of COC/BE with health, forensics, and

bioinformatics application and the potential to be used with other rapid

immunotest strips directed at several other targets. Thus, this report serves as

a general reference and method describing the use of image analysis of lateral

flow rapid test strips.

PMCID: PMC3411535

PMID: 22879741

2090. BMC Med Genet. 2009 Dec 31;10:148. doi: 10.1186/1471-2350-10-148.

Evaluating NAT2PRED for inferring the individual acetylation status from unphased

genotype data.

Sabbagh A(1), Darlu P, Vidaud M.

Author information:

(1)INSERM UMR745, Faculty of Pharmacy, University of Paris Descartes, Paris,

France. audrey.sabbagh@parisdescartes.fr

BACKGROUND: Genetically determined differences in N-acetylation capacity have

proved to be important determinants of both the effectiveness of therapeutic

response and the development of adverse drug reactions and toxicity during drug

treatment. NAT2PRED is a web-server that allows a fast determination of NAT2

acetylation phenotype from genotype data without taking the extra step of

reconstructing haplotypes for each individual (publicly available at

http://nat2pred.rit.albany.edu). However, the classification accuracy of NAT2PRED

needs to be assessed before its application can be advocated at a large scale.

METHODS: The ability of NAT2PRED to classify individuals according to their

acetylation status (slow, intermediate and rapid acetylators) was evaluated in a

worldwide dataset composed of 56 population samples (8,489 individuals) from four

continental regions.

RESULTS: NAT2PRED correctly identified slow acetylators with a sensitivity above

99% for all populations outside sub-Saharan Africa. Nevertheless, NAT2PRED showed

a poor ability to distinguish between intermediate and rapid acetylators, with a

classification error rate reaching up to 10% in the non-African samples.

CONCLUSION: NAT2PRED is an excellent tool to infer the individual acetylation

status from NAT2 genotype data when the main interest is to distinguish slow

acetylators from the others. This should facilitate the determination of the

individual acetylation status in routine clinical practice and lead to better

monitoring of risks associated with cancer and adverse drug reactions.

DOI: 10.1186/1471-2350-10-148

PMCID: PMC2806877

PMID: 20043821 [Indexed for MEDLINE]

2091. BMC Genomics. 2009 Dec 29;10:635. doi: 10.1186/1471-2164-10-635.

3PD: Rapid design of optimal primers for chromosome conformation capture assays.

Fröhler S(1), Dieterich C.

Author information:

(1)Berlin Institute for Medical Systems Biology at the Max Delbrück Center for

Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin-Buch, Germany.

sebastian.froehler@mdc-berlin.de

BACKGROUND: Higher eukaryotes control the expression of their genes by mechanisms

that we are just beginning to understand. A complex layer of control is the

dynamic spatial organization of the nucleus.

RESULTS: We present a bioinformatics solution (3PD) to support the

experimentalist in detecting long-ranging intra or inter chromosomal contacts by

Chromosome conformation capture (3C) assays. 3C assays take a snapshot of

chromosomal contacts by a fixation step and quantify them by PCR. Our

contribution is to rapidly design an optimal primer set for the crucial PCR step.

Our primer design reduces the level of experimental error as primers are highly

similar in terms of physical properties and amplicon length. All 3C primers are

compatible with multiplex PCR reactions. Primer uniqueness is checked genome-wide

with a suitable index structure.

CONCLUSIONS: In summary, our software 3PD facilitates genome-wide primer design

for 3C experiments in a matter of seconds. Our software is available as a web

server at: http://www.pristionchus.org/3CPrimerDesign/.

DOI: 10.1186/1471-2164-10-635

PMCID: PMC2811132

PMID: 20040085 [Indexed for MEDLINE]

2092. BMC Bioinformatics. 2009 Dec 22;10:441. doi: 10.1186/1471-2105-10-441.

The Medicago truncatula gene expression atlas web server.

He J(1), Benedito VA, Wang M, Murray JD, Zhao PX, Tang Y, Udvardi MK.

Author information:

(1)Plant Biology Division, the Samuel Roberts Noble Foundation, Ardmore, OK

73401, USA. jhe@noble.org

BACKGROUND: Legumes (Leguminosae or Fabaceae) play a major role in agriculture.

Transcriptomics studies in the model legume species, Medicago truncatula, are

instrumental in helping to formulate hypotheses about the role of legume genes.

With the rapid growth of publically available Affymetrix GeneChip Medicago Genome

Array GeneChip data from a great range of tissues, cell types, growth conditions,

and stress treatments, the legume research community desires an effective

bioinformatics system to aid efforts to interpret the Medicago genome through

functional genomics. We developed the Medicago truncatula Gene Expression Atlas

(MtGEA) web server for this purpose.

DESCRIPTION: The Medicago truncatula Gene Expression Atlas (MtGEA) web server is

a centralized platform for analyzing the Medicago transcriptome. Currently, the

web server hosts gene expression data from 156 Affymetrix GeneChip(R) Medicago

genome arrays in 64 different experiments, covering a broad range of

developmental and environmental conditions. The server enables flexible,

multifaceted analyses of transcript data and provides a range of additional

information about genes, including different types of annotation and links to the

genome sequence, which help users formulate hypotheses about gene function.

Transcript data can be accessed using Affymetrix probe identification number, DNA

sequence, gene name, functional description in natural language, GO and KEGG

annotation terms, and InterPro domain number. Transcripts can also be discovered

through co-expression or differential expression analysis. Flexible tools to

select a subset of experiments and to visualize and compare expression profiles

of multiple genes have been implemented. Data can be downloaded, in part or full,

in a tabular form compatible with common analytical and visualization software.

The web server will be updated on a regular basis to incorporate new gene

expression data and genome annotation, and is accessible at:

http://bioinfo.noble.org/gene-atlas/.

CONCLUSIONS: The MtGEA web server has a well managed rich data set, and offers

data retrieval and analysis tools provided in the web platform. It's proven to be

a powerful resource for plant biologists to effectively and efficiently identify

Medicago transcripts of interest from a multitude of aspects, formulate

hypothesis about gene function, and overall interpret the Medicago genome from a

systematic point of view.

DOI: 10.1186/1471-2105-10-441

PMCID: PMC2804685

PMID: 20028527 [Indexed for MEDLINE]

2093. BMC Bioinformatics. 2009 Dec 21;10:436. doi: 10.1186/1471-2105-10-436.

PreDisorder: ab initio sequence-based prediction of protein disordered regions.

Deng X(1), Eickholt J, Cheng J.

Author information:

(1)Department of Computer Science, University of Missouri-Columbia, Columbia, MO

65211, USA. xd9d3@mail.missouri.edu

BACKGROUND: Disordered regions are segments of the protein chain which do not

adopt stable structures. Such segments are often of interest because they have a

close relationship with protein expression and functionality. As such, protein

disorder prediction is important for protein structure prediction, structure

determination and function annotation.

RESULTS: This paper presents our protein disorder prediction server, PreDisorder.

It is based on our ab initio prediction method (MULTICOM-CMFR) which, along with

our meta (or consensus) prediction method (MULTICOM), was recently ranked among

the top disorder predictors in the eighth edition of the Critical Assessment of

Techniques for Protein Structure Prediction (CASP8). We systematically

benchmarked PreDisorder along with 26 other protein disorder predictors on the

CASP8 data set and assessed its accuracy using a number of measures. The results

show that it compared favourably with other ab initio methods and its performance

is comparable to that of the best meta and clustering methods.

CONCLUSION: PreDisorder is a fast and reliable server which can be used to

predict protein disordered regions on genomic scale. It is available at

http://casp.rnet.missouri.edu/predisorder.html.

DOI: 10.1186/1471-2105-10-436

PMCID: PMC3087350

PMID: 20025768 [Indexed for MEDLINE]

2094. BMC Bioinformatics. 2009 Dec 19;10:434. doi: 10.1186/1471-2105-10-434.

Identification of ATP binding residues of a protein from its primary sequence.

Chauhan JS(1), Mishra NK, Raghava GP.

Author information:

(1)Institute of Microbial Technology, Chandigarh, India. jagat@imtech.res.in

BACKGROUND: One of the major challenges in post-genomic era is to provide

functional annotations for large number of proteins arising from genome

sequencing projects. The function of many proteins depends on their interaction

with small molecules or ligands. ATP is one such important ligand that plays

critical role as a coenzyme in the functionality of many proteins. There is a

need to develop method for identifying ATP interacting residues in a ATP binding

proteins (ABPs), in order to understand mechanism of protein-ligands interaction.

RESULTS: We have compared the amino acid composition of ATP interacting and

non-interacting regions of proteins and observed that certain residues are

preferred for interaction with ATP. This study describes few models that have

been developed for identifying ATP interacting residues in a protein. All these

models were trained and tested on 168 non-redundant ABPs chains. First we have

developed a Support Vector Machine (SVM) based model using primary sequence of

proteins and obtained maximum MCC 0.33 with accuracy of 66.25%. Secondly, another

SVM based model was developed using position specific scoring matrix (PSSM)

generated by PSI-BLAST. The performance of this model was improved significantly

(MCC 0.5) from the previous one, where only the primary sequence of the proteins

were used.

CONCLUSION: This study demonstrates that it is possible to predict 'ATP

interacting residues' in a protein with moderate accuracy using its sequence. The

evolutionary information is important for the identification of 'ATP interacting

residues', as it provides more information compared to the primary sequence. This

method will be useful for researchers studying ATP-binding proteins. Based on

this study, a web server has been developed for predicting 'ATP interacting

residues' in a protein http://www.imtech.res.in/raghava/atpint/.

DOI: 10.1186/1471-2105-10-434

PMCID: PMC2803200

PMID: 20021687 [Indexed for MEDLINE]

2095. Bioinformatics. 2009 Dec 15;25(24):3319-20. doi: 10.1093/bioinformatics/btp587.

Epub 2009 Oct 14.

ABWGAT: anchor-based whole genome analysis tool.

Das S(1), Vishnoi A, Bhattacharya A.

Author information:

(1)Center for Computational Biology and Bioinformatics, School of Information

Technology, Jawaharlal Nehru University, New Delhi, 110067, India.

dsarbashis@gmail.com

SUMMARY: Large numbers of genomes are being sequenced regularly and the rate will

go up in future due to availability of new genome sequencing techniques. In order

to understand genotype to phenotype relationships, it is necessary to identify

sequence variations at the genomic level. Alignment of a pair of genomes and

parsing the alignment data is an accepted approach for identification of

variations. Though there are a number of tools available for whole-genome

alignment, none of these allows automatic parsing of the alignment and

identification of different kinds of genomic variants with high degree of

sensitivity. Here we present a simple web-based interface for whole genome

comparison named ABWGAT (Anchor-Based Whole Genome Analysis Tool) that is simple

to use. The output is a list of variations such as SNVs, indels, repeat expansion

and inversion.

AVAILABILITY: The web server is freely available to non-commercial users at the

following address http://abwgc.jnu.ac.in/\_sarba. Supplementary data are available

at http://abwgc.jnu.ac.in/\_sarba/cgi-bin/abwgc\_retrival.cgi using job id 524, 526

and 528.

CONTACT: dsarbashis@gmail.com; alok.bhattacharya@gmail.com

DOI: 10.1093/bioinformatics/btp587

PMID: 19828577 [Indexed for MEDLINE]

2096. Bioinformatics. 2009 Dec 15;25(24):3323-4. doi: 10.1093/bioinformatics/btp577.

Epub 2009 Oct 6.

Exon Array Analyzer: a web interface for Affymetrix exon array analysis.

Gellert P(1), Uchida S, Braun T.

Author information:

(1)Max-Planck-Institute for Heart and Lung Research, Parkstrasse 1, 61231 Bad

Nauheim, Germany.

SUMMARY: The Exon Array Analyzer (EAA) is a web server, which provides a

user-friendly interface to identify alternative splicing events analyzed with

Affymetrix Exon Arrays. The EAA implements the Splice Index algorithm to identify

differential expressed exons. The use of various filters allows reduction of the

number of false positive hits. Results are presented with detailed annotation

information and graphics to identify splice events and to facilitate biological

validations. To demonstrate the versatility of the EAA, we analyzed exon arrays

of 11 different murine tissues using sample data provided by Affymetrix

(http://www.affymetrix.com). Data from the heart were compared with other tissues

to identify exons that undergo heart-specific alternatively splicing, resulting

in the identification of 885 differentially expressed probe sets in 649 genes.

AVAILABILITY: The web interface is available at http://EAA.mpi-bn.mpg.de/.

Detailed documentation is available on the EAA web site

(http://EAA.mpi-bn.mpg.de/supp.php) including screen shots, example analyzes and

step by step instructions.

CONTACT: thomas.braun@mpi-bn.mpg.de

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp577

PMID: 19808879 [Indexed for MEDLINE]

2097. Gene. 2009 Dec 15;448(2):207-13. doi: 10.1016/j.gene.2009.07.019. Epub 2009 Aug

3.

Simple and fast classification of non-LTR retrotransposons based on phylogeny of

their RT domain protein sequences.

Kapitonov VV(1), Tempel S, Jurka J.

Author information:

(1)Genetic Information Research Institute, 1925 Landings Dr, Mountain View, CA

94041, USA. vladimir@girinst.org

Rapidly growing number of sequenced genomes requires fast and accurate

computational tools for analysis of different transposable elements (TEs). In

this paper we focus on a rapid and reliable procedure for classification of

autonomous non-LTR retrotransposons based on alignment and clustering of their

reverse transcriptase (RT) domains. Typically, the RT domain protein sequences

encoded by different non-LTR retrotransposons are similar to each other in terms

of significant BLASTP E-values. Therefore, they can be easily detected by the

routine BLASTP searches of genomic DNA sequences coding for proteins similar to

the RT domains of known non-LTR retrotransposons. However, detailed

classification of non-LTR retrotransposons, i.e. their assignment to specific

clades, is a slow and complex procedure that is not formalized or integrated as a

standard set of computational methods and data. Here we describe a tool

(RTclass1) designed for the fast and accurate automated assignment of novel

non-LTR retrotransposons to known or novel clades using phylogenetic analysis of

the RT domain protein sequences. RTclass1 classifies a particular non-LTR

retrotransposon based on its RT domain in less than 10 min on a standard desktop

computer and achieves 99.5% accuracy. RT1class1 works either as a stand-alone

program installed locally or as a web-server that can be accessed distantly by

uploading sequence data through the internet

(http://www.girinst.org/RTphylogeny/RTclass1).

DOI: 10.1016/j.gene.2009.07.019

PMCID: PMC2829327

PMID: 19651192 [Indexed for MEDLINE]

2098. BMC Bioinformatics. 2009 Dec 14;10:416. doi: 10.1186/1471-2105-10-416.

DescFold: a web server for protein fold recognition.

Yan RX(1), Si JN, Wang C, Zhang Z.

Author information:

(1)State Key Laboratory of Agrobiotechnology, College of Biological Sciences,

China Agricultural University, Beijing 100193, China. simpleyrx@163.com

BACKGROUND: Machine learning-based methods have been proven to be powerful in

developing new fold recognition tools. In our previous work [Zhang, Kochhar and

Grigorov (2005) Protein Science, 14: 431-444], a machine learning-based method

called DescFold was established by using Support Vector Machines (SVMs) to

combine the following four descriptors: a profile-sequence-alignment-based

descriptor using Psi-blast e-values and bit scores, a

sequence-profile-alignment-based descriptor using Rps-blast e-values and bit

scores, a descriptor based on secondary structure element alignment (SSEA), and a

descriptor based on the occurrence of PROSITE functional motifs. In this work, we

focus on the improvement of DescFold by incorporating more powerful descriptors

and setting up a user-friendly web server.

RESULTS: In seeking more powerful descriptors, the profile-profile alignment

score generated from the COMPASS algorithm was first considered as a new

descriptor (i.e., PPA). When considering a profile-profile alignment between two

proteins in the context of fold recognition, one protein is regarded as a

template (i.e., its 3D structure is known). Instead of a sequence profile derived

from a Psi-blast search, a structure-seeded profile for the template protein was

generated by searching its structural neighbors with the assistance of the

TM-align structural alignment algorithm. Moreover, the COMPASS algorithm was used

again to derive a profile-structural-profile-alignment-based descriptor (i.e.,

PSPA). We trained and tested the new DescFold in a total of 1,835 highly diverse

proteins extracted from the SCOP 1.73 version. When the PPA and PSPA descriptors

were introduced, the new DescFold boosts the performance of fold recognition

substantially. Using the SCOP\_1.73\_40% dataset as the fold library, the DescFold

web server based on the trained SVM models was further constructed. To provide a

large-scale test for the new DescFold, a stringent test set of 1,866 proteins

were selected from the SCOP 1.75 version. At a less than 5% false positive rate

control, the new DescFold is able to correctly recognize structural homologs at

the fold level for nearly 46% test proteins. Additionally, we also benchmarked

the DescFold method against several well-established fold recognition algorithms

through the LiveBench targets and Lindahl dataset.

CONCLUSIONS: The new DescFold method was intensively benchmarked to have very

competitive performance compared with some well-established fold recognition

methods, suggesting that it can serve as a useful tool to assist in

template-based protein structure prediction. The DescFold server is freely

accessible at http://202.112.170.199/DescFold/index.html.

DOI: 10.1186/1471-2105-10-416

PMCID: PMC2803855

PMID: 20003426 [Indexed for MEDLINE]

2099. BMC Struct Biol. 2009 Dec 14;9:73. doi: 10.1186/1472-6807-9-73.

TIM-Finder: a new method for identifying TIM-barrel proteins.

Si JN(1), Yan RX, Wang C, Zhang Z, Su XD.

Author information:

(1)State Key Laboratory of Agrobiotechnology, College of Biological Sciences,

China Agricultural University, Beijing 100193, China. sijingna@gmail.com

BACKGROUND: The triosephosphate isomerase (TIM)-barrel fold occurs frequently in

the proteomes of different organisms, and the known TIM-barrel proteins have been

found to play diverse functional roles. To accelerate the exploration of the

sequence-structure protein landscape in the TIM-barrel fold, a computational tool

that allows sensitive detection of TIM-barrel proteins is required.

RESULTS: To develop a new TIM-barrel protein identification method in this work,

we consider three descriptors: a sequence-alignment-based descriptor using

PSI-BLAST e-values and bit scores, a descriptor based on secondary structure

element alignment (SSEA), and a descriptor based on the occurrence of PROSITE

functional motifs. With the assistance of Support Vector Machine (SVM), the three

descriptors were combined to obtain a new method with improved performance, which

we call TIM-Finder. When tested on the whole proteome of Bacillus subtilis,

TIM-Finder is able to detect 194 TIM-barrel proteins at a 99% confidence level,

outperforming the PSI-BLAST search as well as one existing fold recognition

method.

CONCLUSIONS: TIM-Finder can serve as a competitive tool for proteome-wide

TIM-barrel protein identification. The TIM-Finder web server is freely accessible

at http://202.112.170.199/TIM-Finder/.

DOI: 10.1186/1472-6807-9-73

PMCID: PMC2803183

PMID: 20003393 [Indexed for MEDLINE]

2100. BMC Bioinformatics. 2009 Dec 13;10:414. doi: 10.1186/1471-2105-10-414.

Modular prediction of protein structural classes from sequences of twilight-zone

identity with predicting sequences.

Mizianty MJ(1), Kurgan L.

Author information:

(1)Department of Electrical and Computer Engineering, University of Alberta,

Edmonton, Canada. mizianty@ualberta.ca

BACKGROUND: Knowledge of structural class is used by numerous methods for

identification of structural/functional characteristics of proteins and could be

used for the detection of remote homologues, particularly for chains that share

twilight-zone similarity. In contrast to existing sequence-based structural class

predictors, which target four major classes and which are designed for high

identity sequences, we predict seven classes from sequences that share

twilight-zone identity with the training sequences.

RESULTS: The proposed MODular Approach to Structural class prediction (MODAS)

method is unique as it allows for selection of any subset of the classes. MODAS

is also the first to utilize a novel, custom-built feature-based sequence

representation that combines evolutionary profiles and predicted secondary

structure. The features quantify information relevant to the definition of the

classes including conservation of residues and arrangement and number of

helix/strand segments. Our comprehensive design considers 8 feature selection

methods and 4 classifiers to develop Support Vector Machine-based classifiers

that are tailored for each of the seven classes. Tests on 5 twilight-zone and 1

high-similarity benchmark datasets and comparison with over two dozens of modern

competing predictors show that MODAS provides the best overall accuracy that

ranges between 80% and 96.7% (83.5% for the twilight-zone datasets), depending on

the dataset. This translates into 19% and 8% error rate reduction when compared

against the best performing competing method on two largest datasets. The

proposed predictor provides accurate predictions at 58% accuracy for membrane

proteins class, which is not considered by majority of existing methods, in spite

that this class accounts for only 2% of the data. Our predictive model is

analyzed to demonstrate how and why the input features are associated with the

corresponding classes.

CONCLUSIONS: The improved predictions stem from the novel features that express

collocation of the secondary structure segments in the protein sequence and that

combine evolutionary and secondary structure information. Our work demonstrates

that conservation and arrangement of the secondary structure segments predicted

along the protein chain can successfully predict structural classes which are

defined based on the spatial arrangement of the secondary structures. A web

server is available at http://biomine.ece.ualberta.ca/MODAS/.

DOI: 10.1186/1471-2105-10-414

PMCID: PMC2805645

PMID: 20003388 [Indexed for MEDLINE]

2101. PLoS One. 2009 Dec 7;4(12):e8116. doi: 10.1371/journal.pone.0008116.

PALM: a paralleled and integrated framework for phylogenetic inference with

automatic likelihood model selectors.

Chen SH(1), Su SY, Lo CZ, Chen KH, Huang TJ, Kuo BH, Lin CY.

Author information:

(1)Institute of Information Science, Academia Sinica, Taipei, Taiwan.

BACKGROUND: Selecting an appropriate substitution model and deriving a tree

topology for a given sequence set are essential in phylogenetic analysis.

However, such time consuming, computationally intensive tasks rely on knowledge

of substitution model theories and related expertise to run through all possible

combinations of several separate programs. To ensure a thorough and efficient

analysis and avert tedious manipulations of various programs, this work presents

an intuitive framework, the phylogenetic reconstruction with automatic likelihood

model selectors (PALM), with convincing, updated algorithms and a best-fit model

selection mechanism for seamless phylogenetic analysis.

METHODOLOGY: As an integrated framework of ClustalW, PhyML, MODELTEST, ProtTest,

and several in-house programs, PALM evaluates the fitness of 56 substitution

models for nucleotide sequences and 112 substitution models for protein sequences

with scores in various criteria. The input for PALM can be either sequences in

FASTA format or a sequence alignment file in PHYLIP format. To accelerate the

computing of maximum likelihood and bootstrapping, this work integrates

MPICH2/PhyML, PalmMonitor and Palm job controller across several machines with

multiple processors and adopts the task parallelism approach. Moreover, an

intuitive and interactive web component, PalmTree, is developed for displaying

and operating the output tree with options of tree rooting, branches swapping,

viewing the branch length values, and viewing bootstrapping score, as well as

removing nodes to restart analysis iteratively.

SIGNIFICANCE: The workflow of PALM is straightforward and coherent. Via a

succinct, user-friendly interface, researchers unfamiliar with phylogenetic

analysis can easily use this server to submit sequences, retrieve the output, and

re-submit a job based on a previous result if some sequences are to be deleted or

added for phylogenetic reconstruction. PALM results in an inference of

phylogenetic relationship not only by vanquishing the computation difficulty of

ML methods but also providing statistic methods for model selection and

bootstrapping. The proposed approach can reduce calculation time, which is

particularly relevant when querying a large data set. PALM can be accessed online

at http://palm.iis.sinica.edu.tw.

DOI: 10.1371/journal.pone.0008116

PMCID: PMC2785425

PMID: 19997614 [Indexed for MEDLINE]

2102. BMC Bioinformatics. 2009 Dec 3;10 Suppl 15:S8. doi: 10.1186/1471-2105-10-S15-S8.

Protein subcellular localization prediction of eukaryotes using a knowledge-based

approach.

Lin HN(1), Chen CT, Sung TY, Ho SY, Hsu WL.

Author information:

(1)Bioinformatics Program, Taiwan International Graduate Program, Academia

Sinica, Taipei, Taiwan, Republic of China. arith@iis.sinica.edu.tw

BACKGROUND: The study of protein subcellular localization (PSL) is important for

elucidating protein functions involved in various cellular processes. However,

determining the localization sites of a protein through wet-lab experiments can

be time-consuming and labor-intensive. Thus, computational approaches become

highly desirable. Most of the PSL prediction systems are established for

single-localized proteins. However, a significant number of eukaryotic proteins

are known to be localized into multiple subcellular organelles. Many studies have

shown that proteins may simultaneously locate or move between different cellular

compartments and be involved in different biological processes with different

roles.

RESULTS: In this study, we propose a knowledge based method, called KnowPredsite,

to predict the localization site(s) of both single-localized and multi-localized

proteins. Based on the local similarity, we can identify the "related sequences"

for prediction. We construct a knowledge base to record the possible sequence

variations for protein sequences. When predicting the localization annotation of

a query protein, we search against the knowledge base and used a scoring

mechanism to determine the predicted sites. We downloaded the dataset from ngLOC,

which consisted of ten distinct subcellular organelles from 1923 species, and

performed ten-fold cross validation experiments to evaluate KnowPred site's

performance. The experiment results show that KnowPred site achieves higher

prediction accuracy than ngLOC and Blast-hit method. For single-localized

proteins, the overall accuracy of KnowPred site is 91.7%. For multi-localized

proteins, the overall accuracy of KnowPred site is 72.1%, which is significantly

higher than that of ngLOC by 12.4%. Notably, half of the proteins in the dataset

that cannot find any Blast hit sequence above a specified threshold can still be

correctly predicted by KnowPred site.

CONCLUSION: KnowPred site demonstrates the power of identifying related sequences

in the knowledge base. The experiment results show that even though the sequence

similarity is low, the local similarity is effective for prediction. Experiment

results show that KnowPred site is a highly accurate prediction method for both

single- and multi-localized proteins. It is worth-mentioning the prediction

process of KnowPred site is transparent and biologically interpretable and it

shows a set of template sequences to generate the prediction result. The KnowPred

site prediction server is available at

http://bio-cluster.iis.sinica.edu.tw/kbloc/.

DOI: 10.1186/1471-2105-10-S15-S8

PMCID: PMC2788359

PMID: 19958518 [Indexed for MEDLINE]

2103. BMC Bioinformatics. 2009 Dec 3;10 Suppl 15:S3. doi: 10.1186/1471-2105-10-S15-S3.

Gevab: a prototype genome variation analysis browsing server.

Kim WY(1), Kim SY, Kim TH, Ahn SM, Byun HN, Kim D, Kim DS, Lee YS, Ghang H, Park

D, Kim BC, Kim C, Lee S, Kim SJ, Bhak J.

Author information:

(1)Korean BioInformation Center (KOBIC), KRIBB, Daejeon, Korea.

kimplove@kribb.re.kr

BACKGROUND: The first Korean individual diploid genome sequence data (KOREF) was

publicized in December 2008.

RESULTS: A Korean genome variation analysis and browsing server (Gevab) was

constructed as a database and web server for the exploration and downloading of

Korean personal genome(s). Information in the Gevab includes SNPs, short indels,

and structural variation (SV) and comparison analysis between the NCBI human

reference and the Korean genome(s). The user can find information on assembled

consensus sequences, sequenced short reads, genetic variations, and relationships

between genotype and phenotypes.

CONCLUSION: This server is openly and publicly available online at

http://koreagenome.org/en/ or directly http://gevab.org.

DOI: 10.1186/1471-2105-10-S15-S3

PMCID: PMC2788354

PMID: 19958513 [Indexed for MEDLINE]

2104. BMC Genomics. 2009 Dec 3;10 Suppl 3:S8. doi: 10.1186/1471-2164-10-S3-S8.

BioBarcode: a general DNA barcoding database and server platform for Asian

biodiversity resources.

Lim J(1), Kim SY, Kim S, Eo HS, Kim CB, Paek WK, Kim W, Bhak J.

Author information:

(1)Korean BioInformation Center, Korea Research Institute of Bioscience and

Biotechnology, Daejeon, 305-806, Korea. jhlim@kribb.re.kr

BACKGROUND: DNA barcoding provides a rapid, accurate, and standardized method for

species-level identification using short DNA sequences. Such a standardized

identification method is useful for mapping all the species on Earth,

particularly when DNA sequencing technology is cheaply available. There are many

nations in Asia with many biodiversity resources that need to be mapped and

registered in databases.

RESULTS: We have built a general DNA barcode data processing system, BioBarcode,

with open source software - which is a general purpose database and server. It

uses mySQL RDBMS 5.0, BLAST2, and Apache httpd server. An exemplary database of

BioBarcode has around 11,300 specimen entries (including GenBank data) and

registers the biological species to map their genetic relationships. The

BioBarcode database contains a chromatogram viewer which improves the performance

in DNA sequence analyses.

CONCLUSION: Asia has a very high degree of biodiversity and the BioBarcode

database server system aims to provide an efficient bioinformatics protocol that

can be freely used by Asian researchers and research organizations interested in

DNA barcoding. The BioBarcode promotes the rapid acquisition of biological

species DNA sequence data that meet global standards by providing specialized

services, and provides useful tools that will make barcoding cheaper and faster

in the biodiversity community such as standardization, depository, management,

and analysis of DNA barcode data. The system can be downloaded upon request, and

an exemplary server has been constructed with which to build an Asian

biodiversity system http://www.asianbarcode.org.

DOI: 10.1186/1471-2164-10-S3-S8

PMCID: PMC2788395

PMID: 19958506 [Indexed for MEDLINE]

2105. BMC Genomics. 2009 Dec 3;10 Suppl 3:S18. doi: 10.1186/1471-2164-10-S3-S18.

PutidaNET: interactome database service and network analysis of Pseudomonas

putida KT2440.

Park SJ(1), Choi JS, Kim BC, Jho SW, Ryu JW, Park D, Lee KA, Bhak J, Kim SI.

Author information:

(1)Korean BioInformation Center (KOBIC), KRIBB, Daejeon 305-806, Korea.

psj420@kribb.re.kr

BACKGROUND: Pseudomonas putida KT2440 (P. putida KT2440) is a highly versatile

saprophytic soil bacterium. It is a certified bio-safety host for transferring

foreign genes. Therefore, the bacterium is used as a model organism for genetic

and physiological studies and for the development of biotechnological

applications. In order to provide a more systematic application of the organism,

we have constructed a protein-protein interaction (PPI) network analysis system

of P. putida KT2440.

RESULTS: PutidaNET is a comprehensive interaction database and server of P.

putida KT2440 which is generated from three protein-protein interaction (PPI)

methods. We used PSIMAP (Protein Structural Interactome MAP), PEIMAP (Protein

Experimental Interactome MAP), and Domain-domain interactions using iPfam.

PutidaNET contains 3,254 proteins, and 82,019 possible interactions consisting of

61,011 (PSIMAP), 4,293 (PEIMAP), and 30,043 (iPfam) interaction pairs except for

self interaction. Also, we performed a case study by integrating a protein

interaction network and experimental 1-DE/MS-MS analysis data P. putida. We found

that 1) major functional modules are involved in various metabolic pathways and

ribosomes, and 2) existing PPI sub-networks that are specific to succinate or

benzoate metabolism are not in the center as predicted.

CONCLUSION: We introduce the PutidaNET which provides predicted interaction

partners and functional analyses such as physicochemical properties, KEGG pathway

assignment, and Gene Ontology mapping of P. putida KT2440 PutidaNET is freely

available at http://sequenceome.kobic.kr/PutidaNET.

DOI: 10.1186/1471-2164-10-S3-S18

PMCID: PMC2788370

PMID: 19958481 [Indexed for MEDLINE]

2106. Biochim Biophys Acta. 2009 Dec;1794(12):1784-94. doi:

10.1016/j.bbapap.2009.08.020. Epub 2009 Aug 28.

3D entropy and moments prediction of enzyme classes and experimental-theoretic

study of peptide fingerprints in Leishmania parasites.

Concu R(1), Dea-Ayuela MA, Perez-Montoto LG, Prado-Prado FJ, Uriarte E,

Bolás-Fernández F, Podda G, Pazos A, Munteanu CR, Ubeira FM, González-Díaz H.

Author information:

(1)Faculty of Pharmacy, Department of Microbiology and Parasitology, University

of Santiago de Compostela, 15782 Santiago de Compostela, Spain.

The number of protein 3D structures without function annotation in Protein Data

Bank (PDB) has been steadily increased. This fact has led in turn to an increment

of demand for theoretical models to give a quick characterization of these

proteins. In this work, we present a new and fast Markov chain model (MCM) to

predict the enzyme classification (EC) number. We used both linear discriminant

analysis (LDA) and/or artificial neural networks (ANN) in order to compare linear

vs. non-linear classifiers. The LDA model found is very simple (three variables)

and at the same time is able to predict the first EC number with an overall

accuracy of 79% for a data set of 4755 proteins (859 enzymes and 3896

non-enzymes) divided into both training and external validation series. In

addition, the best non-linear ANN model is notably more complex but has an

overall accuracy of 98.85%. It is important to emphasize that this method may

help us to predict not only new enzyme proteins but also to select peptide

candidates found on the peptide mass fingerprints (PMFs) of new proteins that may

improve enzyme activity. In order to illustrate the use of the model in this

regard, we first report the 2D electrophoresis (2DE) and MADLI-TOF mass spectra

characterization of the PMF of a new possible malate dehydrogenase sequence from

Leishmania infantum. Next, we used the models to predict the contribution to a

specific enzyme action of 30 peptides found in the PMF of the new protein. We

implemented the present model in a server at portal Bio-AIMS

(http://miaja.tic.udc.es/Bio-AIMS/EnzClassPred.php). This free on-line tool is

based on PHP/HTML/Python and MARCH-INSIDE routines. This combined strategy may be

used to identify and predict peptides of prokaryote and eukaryote parasites and

their hosts as well as other superior organisms, which may be of interest in drug

development or target identification.

DOI: 10.1016/j.bbapap.2009.08.020

PMID: 19716935 [Indexed for MEDLINE]

2107. Bioinformatics. 2009 Dec 1;25(23):3183-4. doi: 10.1093/bioinformatics/btp545.

Epub 2009 Sep 17.

PoreLogo: a new tool to analyse, visualize and compare channels in transmembrane

proteins.

Oliva R(1), Thornton JM, Pellegrini-Calace M.

Author information:

(1)Department of Applied Sciences, University of Parthenope of Naples, Centro

Direzionale Isola C4 80143, Naples, Italy. romina.oliva@uniparthenope.it

The increasing number of available atomic 3D structures of transmembrane channel

proteins represents a valuable resource for better understanding their

structure-function relationships and to eventually predict their selectivity.

Herein, we present PoreLogo, an automatic tool for analysing, visualizing and

comparing the amino acid composition of transmembrane channels and its

conservation across the corresponding protein family.AVAILABILITY: PoreLogo is

accessible as a public web server at

http://www.ebi.ac.uk/thornton-srv/software/PoreLogo/.

DOI: 10.1093/bioinformatics/btp545

PMID: 19762348 [Indexed for MEDLINE]

2108. Genesis. 2009 Dec;47(12):842-6. doi: 10.1002/dvg.20575.

Textpresso site-specific recombinases: A text-mining server for the recombinase

literature including Cre mice and conditional alleles.

Urbanski WM(1), Condie BG.

Author information:

(1)Department of Genetics, Developmental Biology Group, University of Georgia,

Athens, Georgia, USA.

Textpresso Site Specific Recombinases (http://ssrc.genetics.uga.edu/) is a

text-mining web server for searching a database of more than 9,000 full-text

publications. The papers and abstracts in this database represent a wide range of

topics related to site-specific recombinase (SSR) research tools. Included in the

database are most of the papers that report the characterization or use of mouse

strains that express Cre recombinase as well as papers that describe or analyze

mouse lines that carry conditional (floxed) alleles or SSR-activated

transgenes/knockins. The database also includes reports describing SSR-based

cloning methods such as the Gateway or the Creator systems, papers reporting the

development or use of SSR-based tools in systems such as Drosophila, bacteria,

parasites, stem cells, yeast, plants, zebrafish, and Xenopus as well as

publications that describe the biochemistry, genetics, or molecular structure of

the SSRs themselves. Textpresso Site Specific Recombinases is the only

comprehensive text-mining resource available for the literature describing the

biology and technical applications of SSRs.

(c) 2009 Wiley-Liss, Inc.

DOI: 10.1002/dvg.20575

PMCID: PMC4963979

PMID: 19882667 [Indexed for MEDLINE]

2109. Hum Genomics. 2009 Dec;4(2):136-42.

The aldehyde dehydrogenase gene superfamily resource center.

Black W(1), Vasiliou V.

Author information:

(1)Molecular Toxicology and Environmental Health Sciences Program, Department of

Pharmaceutical Sciences, University of Colorado Denver, Aurora, CO 80045, USA.

The website www.aldh.org is a publicly available database for nomenclature and

functional and molecular sequence information for members of the aldehyde

dehydrogenase (ALDH) gene superfamily for animals, plants, fungi and bacteria.

The site has organised gene-specific records. It provides synopses of ALDH gene

records, marries trivial terms to correct nomenclature and links global accession

identifiers with source data. Server-side alignment software characterises the

integrity of each sequence relative to the latest genomic assembly and provides

identifier-specific detail reports, including a graphical presentation of the

transcript's exon-intron structure, its size, coding sequence, genomic strand and

locus. Also included are a summary of substrates, inhibitors and enzyme kinetics.

The site provides reference lists and is designed to facilitate data mining by

interested investigators.

PMCID: PMC3525204

PMID: 20038501 [Indexed for MEDLINE]

2110. J Recept Signal Transduct Res. 2009 Dec;29(6):312-7. doi:

10.3109/10799890903295143.

GRIP: a server for predicting interfaces for GPCR oligomerization.

Nemoto W(1), Fukui K, Toh H.

Author information:

(1)Computational Biology Research Center (CBRC), Advanced Industrial Science and

Technology (AIST), Koto-ku, Tokyo, Japan. w.nemoto@aist.go.jp

G-Protein Coupled Receptors (GPCRs) are one of the most important pharmaceutical

targets. Recent studies have revealed that many GPCRs form homo- and/or

hetero-oligomers. The molecular mechanisms of oligomerization are not fully

understood yet, due to the lack of structural data for GPCR complexes. Therefore,

accurate interface prediction would accelerate investigations of the molecular

mechanisms of oligomerization and signaling via GPCRs. However, interface

prediction for GPCR oligomerization is difficult, because the various GPCR

subtypes often use different structural regions as their interfaces, even when

the subtypes belong to the same subfamily. Previously, we developed a method to

predict the interfaces for GPCR oligomerization, which overcomes the difficulty

described above. We have now launched a web service, named G-protein coupled

Receptors Interaction Partners (GRIP) ( http://grip.cbrc.jp/GRIP/index.html ), to

predict the interfaces for GPCR oligomerization. As far as we know, it is the

only service to predict the interfaces for GPCR oligomerization.

DOI: 10.3109/10799890903295143

PMID: 19888901 [Indexed for MEDLINE]

2111. Magn Reson Chem. 2009 Dec;47 Suppl 1:S118-22. doi: 10.1002/mrc.2486.

Web server suite for complex mixture analysis by covariance NMR.

Zhang F(1), Robinette SL, Bruschweiler-Li L, Brüschweiler R.

Author information:

(1)National High Magnetic Field Laboratory, Florida State University,

Tallahassee, FL 32310, USA.

Elucidation of the chemical composition of biological samples is a main focus of

systems biology and metabolomics. Their comprehensive study requires reliable,

efficient, and automatable methods to identify and quantify the underlying

metabolites. Because nuclear magnetic resonance (NMR) spectroscopy is a rich

source of molecular information, it has a unique potential for this task. Here we

present a suite of public web servers (http://spinportal.magnet.fsu.edu), termed

COLMAR, which facilitates complex mixture analysis by NMR. The COLMAR web portal

presently consists of three servers: COLMAR covariance calculates the covariance

NMR spectrum from an NMR input dataset, such as a TOCSY spectrum; COLMAR DemixC

method decomposes the 2D covariance TOCSY spectrum into a reduced set of

nonredundant 1D cross sections or traces, which belong to individual mixture

components; and COLMAR query screens the traces against a NMR spectral database

to identify individual compounds. Examples are presented that illustrate the

utility of this web server suite for complex mixture analysis.

DOI: 10.1002/mrc.2486

PMCID: PMC2865847

PMID: 19634130 [Indexed for MEDLINE]

2112. Nucleic Acids Res. 2009 Dec;37(22):e152. doi: 10.1093/nar/gkp864.

Selection of hyperfunctional siRNAs with improved potency and specificity.

Wang X(1), Wang X, Varma RK, Beauchamp L, Magdaleno S, Sendera TJ.

Author information:

(1)Department of Radiation Oncology, Washington University School of Medicine, St

Louis, MO 63108, USA. xwang@radonc.wustl.edu

One critical step in RNA interference (RNAi) experiments is to design small

interfering RNAs (siRNAs) that can greatly reduce the expression of the target

transcripts, but not of other unintended targets. Although various statistical

and computational approaches have been attempted, this remains a challenge facing

RNAi researchers. Here, we present a new experimentally validated method for

siRNA design. By analyzing public siRNA data and focusing on hyperfunctional

siRNAs, we identified a set of sequence features as potency selection criteria to

build an siRNA design algorithm with support vector machines. Additional

bioinformatics filters were also included in the algorithm to increase RNAi

specificity by reducing potential sequence cross-hybridization or microRNA-like

effects. Independent validation experiments were performed, which indicated that

the newly designed siRNAs have significantly improved performance, and worked

effectively even at low concentrations. Furthermore, our cell-based studies

demonstrated that the siRNA off-target effects were significantly reduced when

the siRNAs were delivered into cells at the 3 nM concentration compared to 30 nM.

Thus, the capability of our new design program to select highly potent siRNAs

also renders increased RNAi specificity because these siRNAs can be used at a

much lower concentration. The siRNA design web server is available at

http://www5.appliedbiosystems.com/tools/siDesign/.

DOI: 10.1093/nar/gkp864

PMCID: PMC2794195

PMID: 19846596 [Indexed for MEDLINE]

2113. BMC Bioinformatics. 2009 Nov 29;10:391. doi: 10.1186/1471-2105-10-391.

JANE: efficient mapping of prokaryotic ESTs and variable length sequence reads on

related template genomes.

Liang C(1), Schmid A, López-Sánchez MJ, Moya A, Gross R, Bernhardt J, Dandekar T.

Author information:

(1)Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland,

D-97074 Würzburg, Germany. liang@biozentrum.uni-wuerzburg.de

BACKGROUND: ESTs or variable sequence reads can be available in prokaryotic

studies well before a complete genome is known. Use cases include (i)

transcriptome studies or (ii) single cell sequencing of bacteria. Without

suitable software their further analysis and mapping would have to await

finalization of the corresponding genome.

RESULTS: The tool JANE rapidly maps ESTs or variable sequence reads in

prokaryotic sequencing and transcriptome efforts to related template genomes. It

provides an easy-to-use graphics interface for information retrieval and a

toolkit for EST or nucleotide sequence function prediction. Furthermore, we

developed for rapid mapping an enhanced sequence alignment algorithm which

reassembles and evaluates high scoring pairs provided from the BLAST algorithm.

Rapid assembly on and replacement of the template genome by sequence reads or

mapped ESTs is achieved. This is illustrated (i) by data from Staphylococci as

well as from a Blattabacteria sequencing effort, (ii) mapping single cell

sequencing reads is shown for poribacteria to sister phylum representative

Rhodopirellula Baltica SH1. The algorithm has been implemented in a web-server

accessible at http://jane.bioapps.biozentrum.uni-wuerzburg.de.

CONCLUSION: Rapid prokaryotic EST mapping or mapping of sequence reads is

achieved applying JANE even without knowing the cognate genome sequence.

DOI: 10.1186/1471-2105-10-391

PMCID: PMC2789075

PMID: 19943962 [Indexed for MEDLINE]

2114. Bioinform Biol Insights. 2009 Nov 24;1:77-90.

Including Functional Annotations and Extending the Collection of Structural

Classifications of Protein Loops (ArchDB).

Hermoso A(1), Espadaler J, Enrique Querol E, Aviles FX, Sternberg MJ, Oliva B,

Fernandez-Fuentes N.

Author information:

(1)Laboratori de Bioinformàtica, Institut de Biomedicina I Biotecnologia,

Universitat Autònoma de Barcelona, Bellaterra 08193, Catalonia. Spain.

Loops represent an important part of protein structures. The study of loop is

critical for two main reasons: First, loops are often involved in protein

function, stability and folding. Second, despite improvements in experimental and

computational structure prediction methods, modeling the conformation of loops

remains problematic. Here, we present a structural classification of loops,

ArchDB, a mine of information with application in both mentioned fields: loop

structure prediction and function prediction. ArchDB (http://sbi.imim.es/archdb)

is a database of classified protein loop motifs. The current database provides

four different classification sets tailored for different purposes. ArchDB-40, a

loop classification derived from SCOP40, well suited for modeling common loop

motifs. Since features relevant to loop structure or function can be more easily

determined on well-populated clusters, we have developed ArchDB-95, a loop

classification derived from SCOP95. This new classification set shows a ~40%

increase in the number of subclasses, and a large 7-fold increase in the number

of putative structure/function-related subclasses. We also present ArchDB-EC, a

classification of loop motifs from enzymes, and ArchDB-KI, a manually annotated

classification of loop motifs from kinases. Information about ligand contacts and

PDB sites has been included in all classification sets. Improvements in our

classification scheme are described, as well as several new database features,

such as the ability to query by conserved annotations, sequence similarity, or

uploading 3D coordinates of a protein. The lengths of classified loops range

between 0 and 36 residues long. ArchDB offers an exhaustive sampling of loop

structures. Functional information about loops and links with related biological

databases are also provided. All this information and the possibility to

browse/query the database through a web-server outline an useful tool with

application in the comparative study of loops, the analysis of loops involved in

protein function and to obtain templates for loop modeling.

PMCID: PMC2789696

PMID: 20066127

2115. Source Code Biol Med. 2009 Nov 20;4:8. doi: 10.1186/1751-0473-4-8.

HAMSTER: visualizing microarray experiments as a set of minimum spanning trees.

Wan R(1), Kiseleva L, Harada H, Mamitsuka H, Horton P.

Author information:

(1)Bioinformatics Center, Institute for Chemical Research, Kyoto University,

Gokasho, Uji, 611-0011, Japan. r.wan@aist.go.jp.

BACKGROUND: Visualization tools allow researchers to obtain a global view of the

interrelationships between the probes or experiments of a gene expression (e.g.

microarray) data set. Some existing methods include hierarchical clustering and

k-means. In recent years, others have proposed applying minimum spanning trees

(MST) for microarray clustering. Although MST-based clustering is formally

equivalent to the dendrograms produced by hierarchical clustering under certain

conditions; visually they can be quite different.

METHODS: HAMSTER (Helpful Abstraction using Minimum Spanning Trees for Expression

Relations) is an open source system for generating a set of MSTs from the

experiments of a microarray data set. While previous works have generated a

single MST from a data set for data clustering, we recursively merge experiments

and repeat this process to obtain a set of MSTs for data visualization. Depending

on the parameters chosen, each tree is analogous to a snapshot of one step of the

hierarchical clustering process. We scored and ranked these trees using one of

three proposed schemes. HAMSTER is implemented in C++ and makes use of Graphviz

for laying out each MST.

RESULTS: We report on the running time of HAMSTER and demonstrate using data sets

from the NCBI Gene Expression Omnibus (GEO) that the images created by HAMSTER

offer insights that differ from the dendrograms of hierarchical clustering. In

addition to the C++ program which is available as open source, we also provided a

web-based version (HAMSTER+) which allows users to apply our system through a web

browser without any computer programming knowledge.

CONCLUSION: Researchers may find it helpful to include HAMSTER in their

microarray analysis workflow as it can offer insights that differ from

hierarchical clustering. We believe that HAMSTER would be useful for certain

types of gradient data sets (e.g time-series data) and data that indicate

relationships between cells/tissues. Both the source and the web server variant

of HAMSTER are available from http://hamster.cbrc.jp/.

DOI: 10.1186/1751-0473-4-8

PMCID: PMC2784758

PMID: 19925686

2116. BMC Bioinformatics. 2009 Nov 18;10:379. doi: 10.1186/1471-2105-10-379.

SitesIdentify: a protein functional site prediction tool.

Bray T(1), Chan P, Bougouffa S, Greaves R, Doig AJ, Warwicker J.

Author information:

(1)Faculty of Life Sciences, The University of Manchester, Michael Smith

Building, Oxford Road, Manchester M13 9PT, UK. t.bray@postgrad.manchester.ac.uk

BACKGROUND: The rate of protein structures being deposited in the Protein Data

Bank surpasses the capacity to experimentally characterise them and therefore

computational methods to analyse these structures have become increasingly

important. Identifying the region of the protein most likely to be involved in

function is useful in order to gain information about its potential role. There

are many available approaches to predict functional site, but many are not made

available via a publicly-accessible application.

RESULTS: Here we present a functional site prediction tool (SitesIdentify), based

on combining sequence conservation information with geometry-based cleft

identification, that is freely available via a web-server. We have shown that

SitesIdentify compares favourably to other functional site prediction tools in a

comparison of seven methods on a non-redundant set of 237 enzymes with annotated

active sites.

CONCLUSION: SitesIdentify is able to produce comparable accuracy in predicting

functional sites to its closest available counterpart, but in addition achieves

improved accuracy for proteins with few characterised homologues. SitesIdentify

is available via a webserver at

http://www.manchester.ac.uk/bioinformatics/sitesidentify/

DOI: 10.1186/1471-2105-10-379

PMCID: PMC2783165

PMID: 19922660 [Indexed for MEDLINE]

2117. Bioinformation. 2009 Nov 17;4(5):179-81.

JUZBOX: a web server for extracting biomedical words from the protein sequence.

Bobby P(1), Balaji S, Sathyanath V, Eapen SJ.

Author information:

(1)Indian Institute of Spices Research, Calicut, Kerala, India.

The recognition of gene/protein names in literature is one of the pivotal steps

in the processing of biological literatures for information extraction or data

mining. We have compiled a lexicon of biomedical words (conserved patterns/

potential motifs) which has the combination of only 20 alphabets of amino acids.

The remaining 6 letters of the English alphabets (B, J, O, U, X, Z) are treated

as invalid amino acid characters (to our context), We have jumbled the 6 letters

for the sake of usage and convenience and termed as 'JUZBOX' and these characters

were filtered in the biomedical lexicon. Undoubtedly, the generation of

biomedical words from protein sequence using JUZBOX have applications specific

for functional annotation.AVAILABILITY: JUZBOX is available freely at

http://www.spices.res.in/juzbox.

PMCID: PMC2859571

PMID: 20461154

2118. BMC Genomics. 2009 Nov 16;10:529. doi: 10.1186/1471-2164-10-529.

GExplore: a web server for integrated queries of protein domains, gene expression

and mutant phenotypes.

Hutter H(1), Ng MP, Chen N.

Author information:

(1)Department of Biological Sciences, Simon Fraser University, Burnaby, Canada.

hutter@sfu.ca

BACKGROUND: The majority of the genes even in well-studied multi-cellular model

organisms have not been functionally characterized yet. Mining the numerous

genome wide data sets related to protein function to retrieve potential candidate

genes for a particular biological process remains a challenge.

DESCRIPTION: GExplore has been developed to provide a user-friendly database

interface for data mining at the gene expression/protein function level to help

in hypothesis development and experiment design. It supports combinatorial

searches for proteins with certain domains, tissue- or developmental

stage-specific expression patterns, and mutant phenotypes. GExplore operates on a

stand-alone database and has fast response times, which is essential for

exploratory searches. The interface is not only user-friendly, but also modular

so that it accommodates additional data sets in the future.

CONCLUSION: GExplore is an online database for quick mining of data related to

gene and protein function, providing a multi-gene display of data sets related to

the domain composition of proteins as well as expression and phenotype data.

GExplore is publicly available at: http://genome.sfu.ca/gexplore/

DOI: 10.1186/1471-2164-10-529

PMCID: PMC2779824

PMID: 19917126 [Indexed for MEDLINE]

2119. Anal Biochem. 2009 Nov 15;394(2):269-74. doi: 10.1016/j.ab.2009.07.046. Epub 2009

Aug 3.

A top-down approach to enhance the power of predicting human protein subcellular

localization: Hum-mPLoc 2.0.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, 800 Dongchuan Road, Shanghai 200240, China. hbs.hen@jtu.edu.cn

Predicting subcellular localization of human proteins is a challenging problem,

particularly when query proteins may have a multiplex character, i.e.,

simultaneously exist at, or move between, two or more different subcellular

location sites. In a previous study, we developed a predictor called "Hum-mPLoc"

to deal with the multiplex problem for the human protein system. However,

Hum-mPLoc has the following shortcomings. (1) The input of accession number for a

query protein is required in order to obtain a higher expected success rate by

selecting to use the higher-level prediction pathway; but many proteins, such as

synthetic and hypothetical proteins as well as those newly discovered proteins

without being deposited into databanks yet, do not have accession numbers. (2)

Neither functional domain nor sequential evolution information were taken into

account in Hum-mPLoc, and hence its power may be reduced accordingly. In view of

this, a top-down strategy to address these shortcomings has been implemented. The

new predictor thus obtained is called Hum-mPLoc 2.0, where the accession number

for input is no longer needed whatsoever. Moreover, both the functional domain

information and the sequential evolution information have been fused into the

predictor by an ensemble classifier. As a consequence, the prediction power has

been significantly enhanced. The web server of Hum-mPLoc2.0 is freely accessible

at http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/.

DOI: 10.1016/j.ab.2009.07.046

PMID: 19651102 [Indexed for MEDLINE]

2120. Bioinformatics. 2009 Nov 15;25(22):3026-7. doi: 10.1093/bioinformatics/btp523.

Epub 2009 Sep 4.

Saint: a lightweight integration environment for model annotation.

Lister AL(1), Pocock M, Taschuk M, Wipat A.

Author information:

(1)Centre for Integrated Systems Biology of Ageing and Nutrition, Newcastle

University and School of Computing Science, Newcastle University, Newcastle Upon

Tyne, UK. a.l.lister@ncl.ac.uk

Saint is a web application which provides a lightweight annotation integration

environment for quantitative biological models. The system enables modellers to

rapidly mark up models with biological information derived from a range of data

sources.AVAILABILITY AND IMPLEMENTATION: Saint is freely available for use on the

web at http://www.cisban.ac.uk/saint. The web application is implemented in

Google Web Toolkit and Tomcat, with all major browsers supported. The Java source

code is freely available for download at http://saint-annotate.sourceforge.net.

The Saint web server requires an installation of libSBML and has been tested on

Linux (32-bit Ubuntu 8.10 and 9.04).

DOI: 10.1093/bioinformatics/btp523

PMCID: PMC2773255

PMID: 19734151 [Indexed for MEDLINE]

2121. Bioinformatics. 2009 Nov 15;25(22):3031-2. doi: 10.1093/bioinformatics/btp475.

Epub 2009 Aug 4.

PubMed-EX: a web browser extension to enhance PubMed search with text mining

features.

Tsai RT(1), Dai HJ, Lai PT, Huang CH.

Author information:

(1)Department of Computer Science and Engineering, Yuan Ze University, Chung Li,

Taiwan, R.O.C. thtsai@saturn.yzu.edu.tw

PubMed-EX is a browser extension that marks up PubMed search results with

additional text-mining information. PubMed-EX's page mark-up, which includes

section categorization and gene/disease and relation mark-up, can help

researchers to quickly focus on key terms and provide additional information on

them. All text processing is performed server-side, freeing up user

resources.AVAILABILITY: PubMed-EX is freely available at

http://bws.iis.sinica.edu.tw/PubMed-EX and

http://iisr.cse.yzu.edu.tw:8000/PubMed-EX/.

DOI: 10.1093/bioinformatics/btp475

PMID: 19654114 [Indexed for MEDLINE]

2122. J Comput Chem. 2009 Nov 15;30(14):2248-54. doi: 10.1002/jcc.21230.

Multiple classifier integration for the prediction of protein structural classes.

Chen L(1), Lu L, Feng K, Li W, Song J, Zheng L, Yuan Y, Zeng Z, Feng K, Lu W, Cai

Y.

Author information:

(1)Shanghai Key Laboratory of Trustworthy Computing, East China Normal

University, Shanghai 200062, People's Republic of China.

Supervised classifiers, such as artificial neural network, partition trees, and

support vector machines, are often used for the prediction and analysis of

biological data. However, choosing an appropriate classifier is not

straightforward because each classifier has its own strengths and weaknesses, and

each biological dataset has its own characteristics. By integrating many

classifiers together, people can avoid the dilemma of choosing an individual

classifier out of many to achieve an optimized classification results (Rahman et

al., Multiple Classifier Combination for Character Recognition: Revisiting the

Majority Voting System and Its Variation, Springer, Berlin, 2002, 167-178). The

classification algorithms come from Weka (Witten and Frank, Data Mining:

Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco,

2005) (a collection of software tools for machine learning algorithms). By

integrating many predictors (classifiers) together through simple voting, the

correct prediction (classification) rates are 65.21% and 65.63% for a basic

training dataset and an independent test set, respectively. These results are

better than any single machine learning algorithm collected in Weka when exactly

the same data are used. Furthermore, we introduce an integration strategy which

takes care of both classifier weightings and classifier redundancy. A feature

selection strategy, called minimum redundancy maximum relevance (mRMR), is

transferred into algorithm selection to deal with classifier redundancy in this

research, and the weightings are based on the performance of each classifier. The

best classification results are obtained when 11 algorithms are selected by mRMR

method, and integrated together through majority votes with weightings. As a

result, the prediction correct rates are 68.56% and 69.29% for the basic training

dataset and the independent test dataset, respectively. The web-server is

available at http://chemdata.shu.edu.cn/protein\_st/.

2009 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.21230

PMID: 19274708 [Indexed for MEDLINE]

2123. BMC Genomics. 2009 Nov 11;10:517. doi: 10.1186/1471-2164-10-517.

An expression database for roots of the model legume Medicago truncatula under

salt stress.

Li D(1), Su Z, Dong J, Wang T.

Author information:

(1)State Key Laboratory of Agrobiotechnology, College of Biological Sciences,

China Agricultural University, Beijing 100193, PR China.

BACKGROUND: Medicago truncatula is a model legume whose genome is currently being

sequenced by an international consortium. Abiotic stresses such as salt stress

limit plant growth and crop productivity, including those of legumes. We

anticipate that studies on M. truncatula will shed light on other economically

important legumes across the world. Here, we report the development of a database

called MtED that contains gene expression profiles of the roots of M. truncatula

based on time-course salt stress experiments using the Affymetrix Medicago

GeneChip. Our hope is that MtED will provide information to assist in improving

abiotic stress resistance in legumes.

DESCRIPTION: The results of our microarray experiment with roots of M. truncatula

under 180 mM sodium chloride were deposited in the MtED database. Additionally,

sequence and annotation information regarding microarray probe sets were

included. MtED provides functional category analysis based on Gene and GeneBins

Ontology, and other Web-based tools for querying and retrieving query results,

browsing pathways and transcription factor families, showing metabolic maps, and

comparing and visualizing expression profiles. Utilities like mapping probe sets

to genome of M. truncatula and In-Silico PCR were implemented by BLAT software

suite, which were also available through MtED database.

CONCLUSION: MtED was built in the PHP script language and as a MySQL relational

database system on a Linux server. It has an integrated Web interface, which

facilitates ready examination and interpretation of the results of microarray

experiments. It is intended to help in selecting gene markers to improve abiotic

stress resistance in legumes. MtED is available at

http://bioinformatics.cau.edu.cn/MtED/.

DOI: 10.1186/1471-2164-10-517

PMCID: PMC2779821

PMID: 19906315 [Indexed for MEDLINE]

2124. Physiol Genomics. 2009 Nov 6;39(3):172-82. doi:

10.1152/physiolgenomics.90350.2008. Epub 2009 Aug 11.

Effects of atherogenic diet on hepatic gene expression across mouse strains.

Shockley KR(1), Witmer D, Burgess-Herbert SL, Paigen B, Churchill GA.

Author information:

(1)The Jackson Laboratory, Bar Harbor, Maine 04605, USA. shockleykr@niehs.nih.gov

Comment in

Physiol Genomics. 2009 Nov 6;39(3):169-71.

Diets high in fat and cholesterol are associated with increased obesity and

metabolic disease in mice and humans. To study the molecular basis of the

metabolic response to dietary fat, 10 inbred strains of mice were fed atherogenic

high-fat and control low-fat diets. Liver gene expression and whole animal

phenotypes were measured and analyzed in both sexes. The effects of diet, strain,

and sex on gene expression were determined irrespective of complex processes,

such as feedback mechanisms, that could have mediated the genomic responses.

Global gene expression analyses demonstrated that animals of the same strain and

sex have similar transcriptional profiles on a low-fat diet, but strains may show

considerable variability in response to high-fat diet. Functional profiling

indicated that high-fat feeding induced genes in the immune response, indicating

liver damage, and repressed cholesterol biosynthesis. The physiological

significance of the transcriptional changes was confirmed by a correlation

analysis of transcript levels with whole animal phenotypes. The results found

here were used to confirm a previously identified quantitative trait locus on

chromosome 17 identified in males fed a high-fat diet in two crosses, PERA x

DBA/2 and PERA x I/Ln. The gene expression data and phenotype data have been made

publicly available as an online tool for exploring the effects of atherogenic

diet in inbred mouse strains (http://cgd-array.jax.org/DietStrainSurvey).

DOI: 10.1152/physiolgenomics.90350.2008

PMCID: PMC2789673

PMID: 19671657 [Indexed for MEDLINE]

2125. Biosystems. 2009 Nov;98(2):73-9. doi: 10.1016/j.biosystems.2009.06.007. Epub 2009

Jul 5.

Predicting protein subnuclear localization using GO-amino-acid composition

features.

Huang WL(1), Tung CW, Huang HL, Ho SY.

Author information:

(1)Department of Management Information System, Chin Min Institute of Technology,

Miaoli, Taiwan.

The nucleus guides life processes of cells. Many of the nuclear proteins

participating in the life processes tend to concentrate on subnuclear

compartments. The subnuclear localization of nuclear proteins is hence important

for deeply understanding the construction and functions of the nucleus. Recently,

Gene Ontology (GO) annotation has been used for prediction of subnuclear

localization. However, the effective use of GO terms in solving sequence-based

prediction problems remains challenging, especially when query protein sequences

have no accession number or annotated GO term. This study obtains homologies of

query proteins with known accession numbers using BLAST to retrieve GO terms for

sequence-based subnuclear localization prediction. A prediction method PGAC,

which involves mining informative GO terms associated with amino acid composition

features, is proposed to design a support vector machine-based classifier. PGAC

yields 55 informative GO terms with training and test accuracies of 85.7% and

76.3%, respectively, using a data set SNL\_35 (561 proteins in 9 localizations)

with 35% sequence identity. Upon comparison with Nuc-PLoc, which combines

amphiphilic pseudo amino acid composition of a protein with its position-specific

scoring matrix, PGAC using the data set SNL\_80 yields a leave-one-out

cross-validation accuracy of 81.1%, which is better than that of Nuc-PLoc, 67.4%.

Experimental results show that the set of informative GO terms are effective

features for protein subnuclear localization. The prediction server based on PGAC

has been implemented at http://iclab.life.nctu.edu.tw/prolocgac.

DOI: 10.1016/j.biosystems.2009.06.007

PMID: 19583993 [Indexed for MEDLINE]

2126. Int J Med Inform. 2009 Nov;78(11):711-20. doi: 10.1016/j.ijmedinf.2008.09.005.

Epub 2009 Jan 20.

The OpenMRS Implementers Network.

Seebregts CJ(1), Mamlin BW, Biondich PG, Fraser HS, Wolfe BA, Jazayeri D, Allen

C, Miranda J, Baker E, Musinguzi N, Kayiwa D, Fourie C, Lesh N, Kanter A,

Yiannoutsos CT, Bailey C; OpenMRS Implementers Network.

Author information:

(1)Biomedical Informatics Research, eHealth Research and Innovation Platform

Medical Research Council, Cape Town, South Africa. chris.seebregts@mrc.ac.za

OBJECTIVE: OpenMRS (www.openmrs.org) is a configurable open source electronic

medical record application developed and maintained by a large network of open

source developers coordinated by the Regenstrief Institute and Partners in Health

and mainly used for HIV patient and treatment information management in Africa.

Our objective is to develop an open Implementers Network for OpenMRS to provide

regional support for the growing number of OpenMRS implementations in Africa and

to include African developers and implementers in the future growth of OpenMRS.

METHODS: We have developed the OpenMRS Implementers Network using a dedicated

Wiki site and e-mail server. We have also organized annual meetings in South

Africa and regional training courses at African locations where OpenMRS is being

implemented. An OpenMRS Internship program has been initiated and we have started

collaborating with similar networks and projects working in Africa. To evaluate

its potential, OpenMRS was implemented initially at one site in South Africa by a

single implementer using a downloadable OpenMRS application and only the OpenMRS

Implementers Network for support.

RESULTS: The OpenMRS Implementers Network Wiki and list server have grown into

effective means of providing implementation support and forums for exchange of

implementation experiences. The annual OpenMRS Implementers meeting has been held

in South Africa for the past three years and is attracting successively larger

numbers of participants with almost 200 implementers and developers attending the

2008 meeting in Durban, South Africa. Six African developers are presently

registered on the first intake of the OpenMRS Internship program. Successful

collaborations have been started with several African developer groups and

projects initiated to develop interoperability between OpenMRS and various

applications. The South African OpenMRS Implementer group successfully

configured, installed and maintained an integrated HIV/TB OpenMRS application

without significant programming support. Since then, this model has been

replicated in several other African sites. The OpenMRS Implementers Network has

contributed substantially to the growth and sustainability of OpenMRS in Africa

and has become a useful way of including Africans in the development and

implementation of OpenMRS in developing countries. The Network provides valuable

support and enables a basic OpenMRS application to be implemented in the absence

of onsite programmers.

DOI: 10.1016/j.ijmedinf.2008.09.005

PMID: 19157968 [Indexed for MEDLINE]

2127. Mol Divers. 2009 Nov;13(4):475-81. doi: 10.1007/s11030-009-9134-z. Epub 2009 Mar

28.

Using the nonlinear dimensionality reduction method for the prediction of

subcellular localization of Gram-negative bacterial proteins.

Wang T(1), Yang J.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, Shanghai, China. tongwang@sjtu.org

One of the central problems in computational biology is protein function

identification in an automated fashion. A key step to achieve this is predicting

to which subcellular location the protein belongs, since protein localization

correlates closely with its function. A wide variety of methods for protein

subcellular localization prediction have been proposed over recent years. Linear

dimensionality reduction (DR) methods have been introduced to address the

high-dimensionality problem by transforming the representation of protein

sequences. However, this approach is not suitable for some complex biological

systems that have nonlinear characteristics. Herein, we use nonlinear DR methods

such as the kernel DR method to capture the nonlinear characteristics of a

high-dimensional space. Then, the K-nearest-neighbor (K-NN) classifier is

employed to identify the subcellular localization of Gram-negative bacterial

proteins based on their reduced low-dimensional features. Experimental results

thus obtained are quite encouraging, indicating that the applied nonlinear DR

method is effective to deal with this complicated problem of predicting

subcellular localization of Gram-negative bacterial proteins. An online web

server for predicting subcellular location of Gram-negative bacterial proteins is

available at (http://202.120.37.185:8080/).

DOI: 10.1007/s11030-009-9134-z

PMID: 19330461 [Indexed for MEDLINE]

2128. Protein Eng Des Sel. 2009 Nov;22(11):699-705. doi: 10.1093/protein/gzp057. Epub

2009 Sep 22.

GPCR-GIA: a web-server for identifying G-protein coupled receptors and their

families with grey incidence analysis.

Lin WZ(1), Xiao X, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333001,

China.

G-protein-coupled receptors (GPCRs) play fundamental roles in regulating various

physiological processes as well as the activity of virtually all cells. Different

GPCR families are responsible for different functions. With the avalanche of

protein sequences generated in the postgenomic age, it is highly desired to

develop an automated method to address the two problems: given the sequence of a

query protein, can we identify whether it is a GPCR? If it is, what family class

does it belong to? Here, a two-layer ensemble classifier called GPCR-GIA was

proposed by introducing a novel scale called 'grey incident degree'. The overall

success rate by GPCR-GIA in identifying GPCR and non-GPCR was about 95%, and that

in identifying the GPCRs among their nine family classes was about 80%. These

rates were obtained by the jackknife cross-validation tests on the stringent

benchmark data sets where none of the proteins has > or = 50% pairwise sequence

identity to any other in a same class. Moreover, a user-friendly web-server was

established at http://218.65.61.89:8080/bioinfo/GPCR-GIA. For user's convenience,

a step-by-step guide on how to use the GPCR-GIA web server is provided. Generally

speaking, one can get the desired two-level results in around 10 s for a query

protein sequence of 300-400 amino acids; the longer the sequence is, the more

time that is needed.

DOI: 10.1093/protein/gzp057

PMID: 19776029 [Indexed for MEDLINE]

2129. Bioinformation. 2009 Oct 19;4(5):176-8.

CFP: a web-server for constructing sequence-based protein conformational

flexibility profiles.

Kuznetsov IB(1), Rackovsky S.

Author information:

(1)Cancer Research Center, Department of Epidemiology and Biostatistics,

University at Albany, One Discovery Drive, Rensselaer, NY 12144, USA.

ikuznetsov@uamail.albany.edu

Many proteins contain conformationally flexible segments that undergo significant

changes in the backbone conformation or completely lack a well-defined

conformation. Previously, we have developed the generalized local propensity

(GLP), a quantitative sequence-based measure of the protein backbone flexibility.

In this paper, we present the CFP (Conformational Flexibility Profile) web-server

that constructs the GLP flexibility profile for a user-submitted sequence and

uses this profile to identify segments with high backbone flexibility. The

statistical significance of a flexible sequence segment is assessed using the

discrete scan statistics based on the density of flexible residues observed in

this segment.AVAILABILITY: CFP is publicly available at

http://cfp.rit.albany.edu.

PMCID: PMC2859570

PMID: 20461153

2130. BMC Struct Biol. 2009 Oct 19;9:66. doi: 10.1186/1472-6807-9-66.

Machine learning integration for predicting the effect of single amino acid

substitutions on protein stability.

Ozen A(1), Gönen M, Alpaydan E, Haliloğlu T.

Author information:

(1)Department of Chemical Engineering, Polymer Research Center, Boğaziçi

University, Istanbul, Turkey. aozen@prc.boun.edu.tr

BACKGROUND: Computational prediction of protein stability change due to

single-site amino acid substitutions is of interest in protein design and

analysis. We consider the following four ways to improve the performance of the

currently available predictors: (1) We include additional sequence- and

structure-based features, namely, the amino acid substitution likelihoods, the

equilibrium fluctuations of the alpha- and beta-carbon atoms, and the packing

density. (2) By implementing different machine learning integration approaches,

we combine information from different features or representations. (3) We compare

classification vs. regression methods to predict the sign vs. the output of

stability change. (4) We allow a reject option for doubtful cases where the risk

of misclassification is high.

RESULTS: We investigate three different approaches: early, intermediate and late

integration, which respectively combine features, kernels over feature subsets,

and decisions. We perform simulations on two data sets: (1) S1615 is used in

previous studies, (2) S2783 is the updated version (as of July 2, 2009) extracted

also from ProTherm. For S1615 data set, our highest accuracy using both sequence

and structure information is 0.842 on cross-validation and 0.904 on testing using

early integration. Newly added features, namely, local compositional packing and

the mobility extent of the mutated residues, improve accuracy significantly with

intermediate integration. For S2783 data set, we also train regression methods to

estimate not only the sign but also the amount of stability change and apply

risk-based classification to reject when the learner has low confidence and the

loss of misclassification is high. The highest accuracy is 0.835 on

cross-validation and 0.832 on testing using only sequence information. The

percentage of false positives can be decreased to less than 0.005 by rejecting 10

per cent using late integration.

CONCLUSION: We find that in both early and late integration, combining inputs or

decisions is useful in increasing accuracy. Intermediate integration allows

assessing the contributions of individual features by looking at the assigned

weights. Overall accuracy of regression is not better than that of classification

but it has less false positives, especially when combined with the reject option.

The server for stability prediction for three integration approaches and the data

sets are available at http://www.prc.boun.edu.tr/appserv/prc/mlsta.

DOI: 10.1186/1472-6807-9-66

PMCID: PMC2777163

PMID: 19840377 [Indexed for MEDLINE]

2131. Bioinformatics. 2009 Oct 15;25(20):2745-6. doi: 10.1093/bioinformatics/btp518.

Epub 2009 Aug 28.

ANCHOR: web server for predicting protein binding regions in disordered proteins.

Dosztányi Z(1), Mészáros B, Simon I.

Author information:

(1)Institute of Enzymology, Biological Research Center, Hungarian Academy of

Sciences, PO Box 7, H-1518 Budapest, Hungary. zsuzsa@enzim.hu

ANCHOR is a web-based implementation of an original method that takes a single

amino acid sequence as an input and predicts protein binding regions that are

disordered in isolation but can undergo disorder-to-order transition upon

binding. The server incorporates the result of a general disorder prediction

method, IUPred and can carry out simple motif searches as well.AVAILABILITY: The

web server is available at http://anchor.enzim.hu. The program package is freely

available for academic users.

DOI: 10.1093/bioinformatics/btp518

PMCID: PMC2759549

PMID: 19717576 [Indexed for MEDLINE]

2132. Bioinformatics. 2009 Oct 15;25(20):2743-4. doi: 10.1093/bioinformatics/btp512.

Epub 2009 Aug 20.

PiSQRD: a web server for decomposing proteins into quasi-rigid dynamical domains.

Aleksiev T(1), Potestio R, Pontiggia F, Cozzini S, Micheletti C.

Author information:

(1)Scuola Internazionale Superiore di Studi Avanzati and eLab, CNR-INFM

Democritos Simulation Centre, via Beirut 2-4, 34151 Trieste and Italian Institute

of Technology, Genoa, Italy.

SUMMARY: The PiSQRD web resource can be used to subdivide protein structures in

quasi-rigid dynamical domains. The latter are groups of amino acids behaving as

approximately rigid units in the course of protein equilibrium fluctuations. The

PiSQRD server takes as input a biomolecular structure and the desired fraction of

protein internal fluctuations that must be accounted for by the relative

rigid-body motion of the dynamical domains. Next, the lowest energy modes of

fluctuation of the protein (optionally provided by the user) are calculated and

used to identify the rigid subunits. The resulting optimal subdivision is

returned through a web page containing both interactive graphics and detailed

data output.

AVAILABILITY: The PiSQRD web server, which requires Java, is available free of

charge for academic users at http://pisqrd.escience-lab.org.

DOI: 10.1093/bioinformatics/btp512

PMID: 19696046 [Indexed for MEDLINE]

2133. Bioinformatics. 2009 Oct 15;25(20):2625-31. doi: 10.1093/bioinformatics/btp503.

Epub 2009 Aug 19.

TargetMiner: microRNA target prediction with systematic identification of

tissue-specific negative examples.

Bandyopadhyay S(1), Mitra R.

Author information:

(1)Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India.

sanghami@isical.ac.in

MOTIVATION: Prediction of microRNA (miRNA) target mRNAs using machine learning

approaches is an important area of research. However, most of the methods suffer

from either high false positive or false negative rates. One reason for this is

the marked deficiency of negative examples or miRNA non-target pairs. Systematic

identification of non-target mRNAs is still not addressed properly, and

therefore, current machine learning approaches are compelled to rely on

artificially generated negative examples for training.

RESULTS: In this article, we have identified approximately 300 tissue-specific

negative examples using a novel approach that involves expression profiling of

both miRNAs and mRNAs, miRNA-mRNA structural interactions and seed-site

conservation. The newly generated negative examples are validated with pSILAC

dataset, which elucidate the fact that the identified non-targets are indeed

non-targets.These high-throughput tissue-specific negative examples and a set of

experimentally verified positive examples are then used to build a system called

TargetMiner, a support vector machine (SVM)-based classifier. In addition to

assessing the prediction accuracy on cross-validation experiments, TargetMiner

has been validated with a completely independent experimental test dataset. Our

method outperforms 10 existing target prediction algorithms and provides a good

balance between sensitivity and specificity that is not reflected in the existing

methods. We achieve a significantly higher sensitivity and specificity of 69% and

67.8% based on a pool of 90 feature set and 76.5% and 66.1% using a set of 30

selected feature set on the completely independent test dataset. In order to

establish the effectiveness of the systematically generated negative examples,

the SVM is trained using a different set of negative data generated using the

method in Yousef et al. A significantly higher false positive rate (70.6%) is

observed when tested on the independent set, while all other factors are kept the

same. Again, when an existing method (NBmiRTar) is executed with the our proposed

negative data, we observe an improvement in its performance. These clearly

establish the effectiveness of the proposed approach of selecting the negative

examples systematically.

AVAILABILITY: TargetMiner is now available as an online tool at www.isical.ac.in/

approximately bioinfo\_miu

DOI: 10.1093/bioinformatics/btp503

PMID: 19692556 [Indexed for MEDLINE]

2134. BMC Syst Biol. 2009 Oct 2;3:99. doi: 10.1186/1752-0509-3-99.

An editor for pathway drawing and data visualization in the Biopathways

Workbench.

Byrnes RW(1), Cotter D, Maer A, Li J, Nadeau D, Subramaniam S.

Author information:

(1)San Diego Supercomputer Center, University of California San Diego, La Jolla,

CA 92093-0505, USA. rbyrnes@sdsc.edu

BACKGROUND: Pathway models serve as the basis for much of systems biology. They

are often built using programs designed for the purpose. Constructing new models

generally requires simultaneous access to experimental data of diverse types, to

databases of well-characterized biological compounds and molecular intermediates,

and to reference model pathways. However, few if any software applications

provide all such capabilities within a single user interface.

RESULTS: The Pathway Editor is a program written in the Java programming language

that allows de-novo pathway creation and downloading of LIPID MAPS (Lipid

Metabolites and Pathways Strategy) and KEGG lipid metabolic pathways, and of

measured time-dependent changes to lipid components of metabolism. Accessed

through Java Web Start, the program downloads pathways from the LIPID MAPS

Pathway database (Pathway) as well as from the LIPID MAPS web server

http://www.lipidmaps.org. Data arises from metabolomic (lipidomic), microarray,

and protein array experiments performed by the LIPID MAPS consortium of

laboratories and is arranged by experiment. Facility is provided to create,

connect, and annotate nodes and processes on a drawing panel with reference to

database objects and time course data. Node and interaction layout as well as

data display may be configured in pathway diagrams as desired. Users may extend

diagrams, and may also read and write data and non-lipidomic KEGG pathways to and

from files. Pathway diagrams in XML format, containing database identifiers

referencing specific compounds and experiments, can be saved to a local file for

subsequent use. The program is built upon a library of classes, referred to as

the Biopathways Workbench, that convert between different file formats and

database objects. An example of this feature is provided in the form of

read/construct/write access to models in SBML (Systems Biology Markup Language)

contained in the local file system.

CONCLUSION: Inclusion of access to multiple experimental data types and of

pathway diagrams within a single interface, automatic updating through

connectivity to an online database, and a focus on annotation, including

reference to standardized lipid nomenclature as well as common lipid names,

supports the view that the Pathway Editor represents a significant, practicable

contribution to current pathway modeling tools.

DOI: 10.1186/1752-0509-3-99

PMCID: PMC2763869

PMID: 19799790 [Indexed for MEDLINE]

2135. Bioinformatics. 2009 Oct 1;25(19):2566-72. doi: 10.1093/bioinformatics/btp422.

Epub 2009 Jul 9.

aGEM: an integrative system for analyzing spatial-temporal gene-expression

information.

Jiménez-Lozano N(1), Segura J, Macías JR, Vega J, Carazo JM.

Author information:

(1)GN7 of the National Institute for Bioinformatics and Biocomputing Unit of the

National Centre for Biotechnology, Darwin 3, Campus de Cantoblanco, 28049 Madrid,

Spain. natalia@cnb.csic.es

MOTIVATION: The work presented here describes the 'anatomical Gene-Expression

Mapping (aGEM)' Platform, a development conceived to integrate phenotypic

information with the spatial and temporal distributions of genes expressed in the

mouse. The aGEM Platform has been built by extending the Distributed Annotation

System (DAS) protocol, which was originally designed to share genome annotations

over the WWW. DAS is a client-server system in which a single client integrates

information from multiple distributed servers.

RESULTS: The aGEM Platform provides information to answer three main questions.

(i) Which genes are expressed in a given mouse anatomical component? (ii) In

which mouse anatomical structures are a given gene or set of genes expressed? And

(iii) is there any correlation among these findings? Currently, this Platform

includes several well-known mouse resources (EMAGE, GXD and GENSAT), hosting

gene-expression data mostly obtained from in situ techniques together with a

broad set of image-derived annotations.

AVAILABILITY: The Platform is optimized for Firefox 3.0 and it is accessed

through a friendly and intuitive display: http://agem.cnb.csic.es

DOI: 10.1093/bioinformatics/btp422

PMCID: PMC2752607

PMID: 19592395 [Indexed for MEDLINE]

2136. Genomics. 2009 Oct;94(4):284-6. doi: 10.1016/j.ygeno.2009.06.006. Epub 2009 Jun

30.

PGA4genomics for comparative genome assembly based on genetic algorithm

optimization.

Zhao F(1), Hou H, Bao Q, Wu J.

Author information:

(1)Institute of Biomedical Informatics/Zhejiang Provincial Key Laboratory of

Medical Genetics, Wenzhou Medical College, Wenzhou 325035, China. fuz3@psu.edu

New sequencing technologies greatly facilitate the large-scale bacterial genome

sequencing by reducing cost. However, a considerable bottleneck is in the

finishing phase, where dozens to hundreds of gaps need to be closed. In this

study, we constructed a web server (PGA4genomics) to help users automate gap

closing based on comparative genomic syntenies. Extensive evaluations showed that

it significantly outperforms previous methods and can produce highly accurate

layout result, especially when assembling genomes that are only moderately

related. The availability of such a platform would greatly benefit the research

community working on bacterial genomics. PGA4genomics can be accessed at two

mirror sites http://centre.bioinformatics.zj.cn:8080/pga or

http://59.79.168.90:8080/pga.

DOI: 10.1016/j.ygeno.2009.06.006

PMID: 19573591 [Indexed for MEDLINE]

2137. Int J Syst Evol Microbiol. 2009 Oct;59(Pt 10):2582-93. doi:

10.1099/ijs.0.010249-0. Epub 2009 Jul 21.

Construction of an interactive online phytoplasma classification tool,

iPhyClassifier, and its application in analysis of the peach X-disease

phytoplasma group (16SrIII).

Zhao Y(1), Wei W, Lee IM, Shao J, Suo X, Davis RE.

Author information:

(1)Molecular Plant Pathology Laboratory, USDA-Agricultural Research Service,

Beltsville, MD 20705, USA. yan.zhao@ars.usda.gov

Phytoplasmas, the causal agents of numerous plant diseases, are

insect-vector-transmitted, cell-wall-less bacteria descended from ancestral

low-G+C-content Gram-positive bacteria in the Bacillus-Clostridium group. Despite

their monophyletic origin, widely divergent phytoplasma lineages have evolved in

adaptation to specific ecological niches. Classification and taxonomic assignment

of phytoplasmas have been based primarily on molecular analysis of 16S rRNA gene

sequences because of the inaccessibility of measurable phenotypic characters

suitable for conventional microbial characterization. In the present study, an

interactive online tool, iPhyClassifier, was developed to expand the efficacy and

capacity of the current 16S rRNA gene sequence-based phytoplasma classification

system. iPhyClassifier performs sequence similarity analysis, simulates

laboratory restriction enzyme digestions and subsequent gel electrophoresis and

generates virtual restriction fragment length polymorphism (RFLP) profiles. Based

on calculated RFLP pattern similarity coefficients and overall sequence

similarity scores, iPhyClassifier makes instant suggestions on tentative

phytoplasma 16Sr group/subgroup classification status and 'Candidatus

Phytoplasma' species assignment. Using iPhyClassifier, we revised and updated the

classification of strains affiliated with the peach X-disease phytoplasma group.

The online tool can be accessed at

http://www.ba.ars.usda.gov/data/mppl/iPhyClassifier.html.

DOI: 10.1099/ijs.0.010249-0

PMCID: PMC2884932

PMID: 19622670 [Indexed for MEDLINE]

2138. Proteomics. 2009 Oct;9(20):4669-73. doi: 10.1002/pmic.200900273.

Yeast proteome map (last update).

Perrot M(1), Moes S, Massoni A, Jenoe P, Boucherie H.

Author information:

(1)IBGC, UMR CNRS 5095 Bordeaux, France.

The identification of proteins separated on 2-D gels is essential to exploit the

full potential of 2-D gel electrophoresis for proteomic investigations. For this

purpose we have undertaken the systematic identification of Saccharomyces

cerevisiae proteins separated on 2-D gels. We report here the identification by

mass spectrometry of 100 novel yeast protein spots that have so far not been

tackled due to their scarcity on our standard 2-D gels. These identifications

extend the number of protein spots identified on our yeast 2-D proteome map to

716. They correspond to 485 unique proteins. Among these, 154 were resolved into

several isoforms. The present data set can now be expanded to report for the

first time a map of 363 protein isoforms that significantly deepens our knowledge

of the yeast proteome. The reference map and a list of all identified proteins

can be accessed on the Yeast Protein Map server (www.ibgc.u-bordeaux2.fr/YPM).

DOI: 10.1002/pmic.200900273

PMID: 19743426 [Indexed for MEDLINE]

2139. BMC Bioinformatics. 2009 Sep 28;10:314. doi: 10.1186/1471-2105-10-314.

MetaTM - a consensus method for transmembrane protein topology prediction.

Klammer M(1), Messina DN, Schmitt T, Sonnhammer EL.

Author information:

(1)Stockholm Bioinformatics Centre, Albanova, Stockholm University, 10691

Stockholm, Sweden. klammer@sbc.su.se

BACKGROUND: Transmembrane (TM) proteins are proteins that span a biological

membrane one or more times. As their 3-D structures are hard to determine,

experiments focus on identifying their topology (i. e. which parts of the amino

acid sequence are buried in the membrane and which are located on either side of

the membrane), but only a few topologies are known. Consequently, various

computational TM topology predictors have been developed, but their accuracies

are far from perfect. The prediction quality can be improved by applying a

consensus approach, which combines results of several predictors to yield a more

reliable result.

RESULTS: A novel TM consensus method, named MetaTM, is proposed in this work.

MetaTM is based on support vector machine models and combines the results of six

TM topology predictors and two signal peptide predictors. On a large data set

comprising 1460 sequences of TM proteins with known topologies and 2362 globular

protein sequences it correctly predicts 86.7% of all topologies.

CONCLUSION: Combining several TM predictors in a consensus prediction framework

improves overall accuracy compared to any of the individual methods. Our proposed

SVM-based system also has higher accuracy than a previous consensus predictor.

MetaTM is made available both as downloadable source code and as DAS server at

http://MetaTM.sbc.su.se.

DOI: 10.1186/1471-2105-10-314

PMCID: PMC2761906

PMID: 19785723 [Indexed for MEDLINE]

2140. J Med Chem. 2009 Sep 24;52(18):5673-84. doi: 10.1021/jm8016464.

Structural artifacts in protein-ligand X-ray structures: implications for the

development of docking scoring functions.

Søndergaard CR(1), Garrett AE, Carstensen T, Pollastri G, Nielsen JE.

Author information:

(1)School of Biomolecular and Biomedical Science, Centre for Synthesis and

Chemical Biology, UCD Conway Institute, University College Dublin, Belfield,

Dublin 4, Ireland.

The development of docking scoring functions requires high-resolution 3D

structures of protein-ligand complexes for which the binding affinity of the

ligand has been measured experimentally. Protein-ligand binding affinities are

measured in solution experiments, and high resolution protein-ligand structures

can be determined only by X-ray crystallography. Protein-ligand scoring functions

must therefore reproduce solution binding energies using analyses of proteins in

a crystal environment. We present an analysis of the prevalence of

crystal-induced artifacts and water-mediated contacts in protein-ligand complexes

and demonstrate the effect that these can have on the performance of

protein-ligand scoring functions. We find 36% of ligands in the PDBBind 2007

refined data set to be influenced by crystal contacts and find the performance of

a scoring function to be affected by these. A Web server for detecting crystal

contacts in protein-ligand complexes is available at

http://enzyme.ucd.ie/LIGCRYST .

DOI: 10.1021/jm8016464

PMID: 19711919 [Indexed for MEDLINE]

2141. BMC Bioinformatics. 2009 Sep 23;10:309. doi: 10.1186/1471-2105-10-309.

SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services.

Gessler DD(1), Schiltz GS, May GD, Avraham S, Town CD, Grant D, Nelson RT.

Author information:

(1)University of Arizona, Tucson, AZ 85721, USA. dgessler@iplantcollaborative.org

BACKGROUND: SSWAP (Simple Semantic Web Architecture and Protocol; pronounced

"swap") is an architecture, protocol, and platform for using reasoning to

semantically integrate heterogeneous disparate data and services on the web.

SSWAP was developed as a hybrid semantic web services technology to overcome

limitations found in both pure web service technologies and pure semantic web

technologies.

RESULTS: There are currently over 2400 resources published in SSWAP.

Approximately two dozen are custom-written services for QTL (Quantitative Trait

Loci) and mapping data for legumes and grasses (grains). The remaining are

wrappers to Nucleic Acids Research Database and Web Server entries. As an

architecture, SSWAP establishes how clients (users of data, services, and

ontologies), providers (suppliers of data, services, and ontologies), and

discovery servers (semantic search engines) interact to allow for the

description, querying, discovery, invocation, and response of semantic web

services. As a protocol, SSWAP provides the vocabulary and semantics to allow

clients, providers, and discovery servers to engage in semantic web services. The

protocol is based on the W3C-sanctioned first-order description logic language

OWL DL. As an open source platform, a discovery server running at

http://sswap.info (as in to "swap info") uses the description logic reasoner

Pellet to integrate semantic resources. The platform hosts an interactive guide

to the protocol at http://sswap.info/protocol.jsp, developer tools at

http://sswap.info/developer.jsp, and a portal to third-party ontologies at

http://sswapmeet.sswap.info (a "swap meet").

CONCLUSION: SSWAP addresses the three basic requirements of a semantic web

services architecture (i.e., a common syntax, shared semantic, and semantic

discovery) while addressing three technology limitations common in distributed

service systems: i.e., i) the fatal mutability of traditional interfaces, ii) the

rigidity and fragility of static subsumption hierarchies, and iii) the

confounding of content, structure, and presentation. SSWAP is novel by

establishing the concept of a canonical yet mutable OWL DL graph that allows data

and service providers to describe their resources, to allow discovery servers to

offer semantically rich search engines, to allow clients to discover and invoke

those resources, and to allow providers to respond with semantically tagged data.

SSWAP allows for a mix-and-match of terms from both new and legacy third-party

ontologies in these graphs.

DOI: 10.1186/1471-2105-10-309

PMCID: PMC2761904

PMID: 19775460 [Indexed for MEDLINE]

2142. BMC Bioinformatics. 2009 Sep 18;10:295. doi: 10.1186/1471-2105-10-295.

Accurate microRNA target prediction correlates with protein repression levels.

Maragkakis M(1), Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, Giannopoulos

G, Goumas G, Koukis E, Kourtis K, Simossis VA, Sethupathy P, Vergoulis T, Koziris

N, Sellis T, Tsanakas P, Hatzigeorgiou AG.

Author information:

(1)Institute of Molecular Oncology, Biomedical Sciences Research Center Alexander

Fleming, Vari, Greece. maragkakis@fleming.gr

BACKGROUND: MicroRNAs are small endogenously expressed non-coding RNA molecules

that regulate target gene expression through translation repression or messenger

RNA degradation. MicroRNA regulation is performed through pairing of the microRNA

to sites in the messenger RNA of protein coding genes. Since experimental

identification of miRNA target genes poses difficulties, computational microRNA

target prediction is one of the key means in deciphering the role of microRNAs in

development and disease.

RESULTS: DIANA-microT 3.0 is an algorithm for microRNA target prediction which is

based on several parameters calculated individually for each microRNA and

combines conserved and non-conserved microRNA recognition elements into a final

prediction score, which correlates with protein production fold change.

Specifically, for each predicted interaction the program reports a signal to

noise ratio and a precision score which can be used as an indication of the false

positive rate of the prediction.

CONCLUSION: Recently, several computational target prediction programs were

benchmarked based on a set of microRNA target genes identified by the pSILAC

method. In this assessment DIANA-microT 3.0 was found to achieve the highest

precision among the most widely used microRNA target prediction programs reaching

approximately 66%. The DIANA-microT 3.0 prediction results are available online

in a user friendly web server at http://www.microrna.gr/microT.

DOI: 10.1186/1471-2105-10-295

PMCID: PMC2752464

PMID: 19765283 [Indexed for MEDLINE]

2143. Source Code Biol Med. 2009 Sep 17;4:6. doi: 10.1186/1751-0473-4-6.

A web server for interactive and zoomable Chaos Game Representation images.

Arakawa K(1), Oshita K, Tomita M.

Author information:

(1)Institute for Advanced Biosciences, Keio University, Fujisawa, 252-8520,

Japan. gaou@sfc.keio.ac.jp.

Chaos Game Representation (CGR) is a generalized scale-independent Markov

transition table, which is useful for the visualization and comparative study of

genomic signature, or for the study of characteristic sequence motifs. However,

in order to fully utilize the scale-independent properties of CGR, it should be

accessible through scale-independent user interface instead of static images.

Here we describe a web server and Perl library for generating zoomable CGR images

utilizing Google Maps API, which is also easily searchable for specific motifs.

The web server is freely accessible at http://www.g-language.org/wiki/cgr/, and

the Perl library as well as the source code is distributed with the G-language

Genome Analysis Environment under GNU General Public License.

DOI: 10.1186/1751-0473-4-6

PMCID: PMC2753581

PMID: 19761591

2144. Bioinformatics. 2009 Sep 15;25(18):2425-9. doi: 10.1093/bioinformatics/btp430.

Epub 2009 Jul 14.

WebArrayDB: cross-platform microarray data analysis and public data repository.

Xia XQ(1), McClelland M, Porwollik S, Song W, Cong X, Wang Y.

Author information:

(1)Vaccine Research Institute San Diego, San Diego, CA 92121, USA.

xqxia70@gmail.com

MOTIVATION: Cross-platform microarray analysis is an increasingly important

research tool, but researchers still lack open source tools for storing,

integrating and analyzing large amounts of microarray data obtained from

different array platforms.

RESULTS: An open source integrated microarray database and analysis suite,

WebArrayDB (http://www.webarraydb.org), has been developed that features

convenient uploading of data for storage in a MIAME (Minimal Information about a

Microarray Experiment) compliant fashion, and allows data to be mined with a

large variety of R-based tools, including data analysis across multiple

platforms. Different methods for probe alignment, normalization and statistical

analysis are included to account for systematic bias. Student's t-test, moderated

t-tests, non-parametric tests and analysis of variance or covariance

(ANOVA/ANCOVA) are among the choices of algorithms for differential analysis of

data. Users also have the flexibility to define new factors and create new

analysis models to fit complex experimental designs. All data can be queried or

browsed through a web browser. The computations can be performed in parallel on

symmetric multiprocessing (SMP) systems or Linux clusters.

AVAILABILITY: The software package is available for the use on a public web

server (http://www.webarraydb.org) or can be downloaded. are available at

Bioinformatics online.

DOI: 10.1093/bioinformatics/btp430

PMCID: PMC2735672

PMID: 19602526 [Indexed for MEDLINE]

2145. BMC Bioinformatics. 2009 Sep 14;10:287. doi: 10.1186/1471-2105-10-287.

Epitopia: a web-server for predicting B-cell epitopes.

Rubinstein ND(1), Mayrose I, Martz E, Pupko T.

Author information:

(1)Department of Cell Research and Immunology, George S Wise Faculty of Life

Sciences, Tel Aviv University, Tel Aviv 69978, Israel. rubi@post.tau.ac.il

BACKGROUND: Detecting candidate B-cell epitopes in a protein is a basic and

fundamental step in many immunological applications. Due to the impracticality of

experimental approaches to systematically scan the entire protein, a

computational tool that predicts the most probable epitope regions is desirable.

RESULTS: The Epitopia server is a web-based tool that aims to predict immunogenic

regions in either a protein three-dimensional structure or a linear sequence.

Epitopia implements a machine-learning algorithm that was trained to discern

antigenic features within a given protein. The Epitopia algorithm has been

compared to other available epitope prediction tools and was found to have higher

predictive power. A special emphasis was put on the development of a

user-friendly graphical interface for displaying the results.

CONCLUSION: Epitopia is a user-friendly web-server that predicts immunogenic

regions for both a protein structure and a protein sequence. Its accuracy and

functionality make it a highly useful tool. Epitopia is available at

http://epitopia.tau.ac.il and includes extensive explanations and example

predictions.

DOI: 10.1186/1471-2105-10-287

PMCID: PMC2751785

PMID: 19751513 [Indexed for MEDLINE]

2146. Genomics Proteomics Bioinformatics. 2009 Sep;7(3):138-42. doi:

10.1016/S1672-0229(08)60042-X.

PBOND: web server for the prediction of proline and non-proline cis/trans

isomerization.

Exarchos KP(1), Exarchos TP, Papaloukas C, Troganis AN, Fotiadis DI.

Author information:

(1)Unit of Medical Technology and Intelligent Information Systems, Department of

Computer Science, University of loannina, loannina 45110, Greece.

PBOND is a web server that predicts the conformation of the peptide bond between

any two amino acids. PBOND classifies the peptide bonds into one out of four

classes, namely cis imide (cis-Pro), cis amide (cis-nonPro), trans imide

(trans-Pro) and trans amide (trans-nonPro). Moreover, for every prediction a

reliability index is computed. The underlying structure of the server consists of

three stages: (1) feature extraction, (2) feature selection and (3) peptide bond

classification. PBOND can handle both single sequences as well as multiple

sequences for batch processing. The predictions can either be directly downloaded

from the web site or returned via e-mail. The PBOND web server is freely

available at http://195.251.198.21/pbond.html.

DOI: 10.1016/S1672-0229(08)60042-X

PMCID: PMC5054403

PMID: 19944386 [Indexed for MEDLINE]

2147. J Proteome Res. 2009 Sep;8(9):4362-71. doi: 10.1021/pr900204r.

The bologna annotation resource: a non hierarchical method for the functional and

structural annotation of protein sequences relying on a comparative large-scale

genome analysis.

Bartoli L(1), Montanucci L, Fronza R, Martelli PL, Fariselli P, Carota L, Donvito

G, Maggi GP, Casadio R.

Author information:

(1)Biocomputing Group, CIRB/Dept of Biology, University of Bologna, Italy.

Protein sequence annotation is a major challenge in the postgenomic era. Thanks

to the availability of complete genomes and proteomes, protein annotation has

recently taken invaluable advantage from cross-genome comparisons. In this work,

we describe a new non hierarchical clustering procedure characterized by a

stringent metric which ensures a reliable transfer of function between related

proteins even in the case of multidomain and distantly related proteins. The

method takes advantage of the comparative analysis of 599 completely sequenced

genomes, both from prokaryotes and eukaryotes, and of a GO and PDB/SCOP mapping

over the clusters. A statistical validation of our method demonstrates that our

clustering technique captures the essential information shared between homologous

and distantly related protein sequences. By this, uncharacterized proteins can be

safely annotated by inheriting the annotation of the cluster. We validate our

method by blindly annotating other 201 genomes and finally we develop BAR (the

Bologna Annotation Resource), a prediction server for protein functional

annotation based on a total of 800 genomes (publicly available at

http://microserf.biocomp.unibo.it/bar/).

DOI: 10.1021/pr900204r

PMID: 19552451 [Indexed for MEDLINE]

2148. Proteins. 2009 Sep;76(4):930-45. doi: 10.1002/prot.22401.

Building and assessing atomic models of proteins from structural templates:

learning and benchmarks.

Vallat BK(1), Pillardy J, Májek P, Meller J, Blom T, Cao B, Elber R.

Author information:

(1)Department of Chemistry and Biochemistry, Institute of Computational

Engineering and Sciences, University of Texas at Austin, Austin, Texas 78712,

USA.

One approach to predict a protein fold from a sequence (a target) is based on

structures of related proteins that are used as templates. We present an

algorithm that examines a set of candidates for templates, builds from each of

the templates an atomically detailed model, and ranks the models. The algorithm

performs a hierarchical selection of the best model using a diverse set of

signals. After a quick and suboptimal screening of template candidates from the

protein data bank, the current method fine-tunes the selection to a few models.

More detailed signals test the compatibility of the sequence and the proposed

structures, and are merged to give a global fitness measure using linear

programming. This algorithm is a component of the prediction server LOOPP

(http://www.loopp.org). Large-scale training and tests sets were designed and are

presented. Recent results of the LOOPP server in CASP8 are discussed.

Copyright 2009 Wiley-Liss, Inc.

DOI: 10.1002/prot.22401

PMCID: PMC2719020

PMID: 19326457 [Indexed for MEDLINE]

2149. BMC Bioinformatics. 2009 Aug 27;10 Suppl 8:S5. doi: 10.1186/1471-2105-10-S8-S5.

An integrated approach to the interpretation of single amino acid polymorphisms

within the framework of CATH and Gene3D.

Izarzugaza JM(1), Baresic A, McMillan LE, Yeats C, Clegg AB, Orengo CA, Martin

AC, Valencia A.

Author information:

(1)Institute of Structural and Molecular Biology, University College London, UK.

jmgonzalez@cnio.es

BACKGROUND: The phenotypic effects of sequence variations in protein-coding

regions come about primarily via their effects on the resulting structures, for

example by disrupting active sites or affecting structural stability. In order

better to understand the mechanisms behind known mutant phenotypes, and predict

the effects of novel variations, biologists need tools to gauge the impacts of

DNA mutations in terms of their structural manifestation. Although many mutations

occur within domains whose structure has been solved, many more occur within

genes whose protein products have not been structurally characterized.

RESULTS: Here we present 3DSim (3D Structural Implication of Mutations), a

database and web application facilitating the localization and visualization of

single amino acid polymorphisms (SAAPs) mapped to protein structures even where

the structure of the protein of interest is unknown. The server displays

information on 6514 point mutations, 4865 of them known to be associated with

disease. These polymorphisms are drawn from SAAPdb, which aggregates data from

various sources including dbSNP and several pathogenic mutation databases. While

the SAAPdb interface displays mutations on known structures, 3DSim projects

mutations onto known sequence domains in Gene3D. This resource contains sequences

annotated with domains predicted to belong to structural families in the CATH

database. Mappings between domain sequences in Gene3D and known structures in

CATH are obtained using a MUSCLE alignment. 1210 three-dimensional structures

corresponding to CATH structural domains are currently included in 3DSim; these

domains are distributed across 396 CATH superfamilies, and provide a

comprehensive overview of the distribution of mutations in structural space.

CONCLUSION: The server is publicly available at http://3DSim.bioinfo.cnio.es/. In

addition, the database containing the mapping between SAAPdb, Gene3D and CATH is

available on request and most of the functionality is available through

programmatic web service access.

DOI: 10.1186/1471-2105-10-S8-S5

PMCID: PMC2745587

PMID: 19758469 [Indexed for MEDLINE]

2150. J Integr Bioinform. 2009 Aug 23;6(1):108. doi: 10.2390/biecoll-jib-2009-108.

Goober: a fully integrated and user-friendly microarray data management and

analysis solution for core labs and bench biologists.

Luo W(1), Gudipati M, Jung K, Chen M, Marschke KB.

Author information:

(1)Ligand Pharmaceuticals INC, San Diego, CA 92121, USA. wluo@ligand.com

Despite the large number of software tools developed to address different areas

of microarray data analysis, very few offer an all-in-one solution with little

learning curve. For microarray core labs, there are even fewer software packages

available to help with their routine but critical tasks, such as data quality

control (QC) and inventory management. We have developed a simple-to-use web

portal to allow bench biologists to analyze and query complicated microarray data

and related biological pathways without prior training. Both experiment-based and

gene-based analysis can be easily performed, even for the first-time user,

through the intuitive multi-layer design and interactive graphic links. While

being friendly to inexperienced users, most parameters in Goober can be easily

adjusted via drop-down menus to allow advanced users to tailor their needs and

perform more complicated analysis. Moreover, we have integrated graphic pathway

analysis into the website to help users examine microarray data within the

relevant biological content. Goober also contains features that cover most of the

common tasks in microarray core labs, such as real time array QC, data loading,

array usage and inventory tracking. Overall, Goober is a complete microarray

solution to help biologists instantly discover valuable information from a

microarray experiment and enhance the quality and productivity of microarray core

labs. The whole package is freely available at

http://sourceforge.net/projects/goober. A demo web server is available at

http://www.goober-array.org.

DOI: 10.2390/biecoll-jib-2009-108

PMID: 20134074 [Indexed for MEDLINE]

2151. Bioinformatics. 2009 Aug 15;25(16):2076-7. doi: 10.1093/bioinformatics/btp346.

Epub 2009 Jun 3.

PROCAIN server for remote protein sequence similarity search.

Wang Y(1), Sadreyev RI, Grishin NV.

Author information:

(1)Biomedical Engineering Program, University of Texas Southwestern Medical

Center, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd,

Dallas, TX 75390-9050, USA.

Sensitive and accurate detection of distant protein homology is essential for the

studies of protein structure, function and evolution. We recently developed

PROCAIN, a method that is based on sequence profile comparison and involves the

analysis of four signals--similarities of residue content at the profile

positions combined with three types of assisting information: sequence motifs,

residue conservation and predicted secondary structure. Here we present the

PROCAIN web server that allows the user to submit a query sequence or multiple

sequence alignment and perform the search in a profile database of choice. The

output is structured similar to that of BLAST, with the list of detected homologs

sorted by E-value and followed by profile-profile alignments. The front page

allows the user to adjust multiple options of input processing and output

formatting, as well as search settings, including the relative weights assigned

to the three types of assisting information.AVAILABILITY:

http://prodata.swmed.edu/procain/.

DOI: 10.1093/bioinformatics/btp346

PMCID: PMC2723001

PMID: 19497935 [Indexed for MEDLINE]

2152. Proteins. 2009 Aug 15;76(3):718-30. doi: 10.1002/prot.22384.

An all-atom knowledge-based energy function for protein-DNA threading, docking

decoy discrimination, and prediction of transcription-factor binding profiles.

Xu B(1), Yang Y, Liang H, Zhou Y.

Author information:

(1)Department of Polymer Science and Engineering, University of Science and

Technology of China, Hefei, Anhui, China.

How to make an accurate representation of protein-DNA interaction by an energy

function is a long-standing unsolved problem in structural biology. Here, we

modified a statistical potential based on the distance-scaled, finite ideal-gas

reference state so that it is optimized for protein-DNA interactions. The changes

include a volume-fraction correction to account for unmixable atom types in

proteins and DNA in addition to the usage of a low-count correction,

residue/base-specific atom types, and a shorter cutoff distance for protein-DNA

interactions. The new statistical energy functions are tested in threading and

docking decoy discriminations and prediction of protein-DNA binding affinities

and transcription-factor binding profiles. The results indicate that new proposed

energy functions are among the best in existing energy functions for protein-DNA

interactions. The new energy functions are available as a web-server called DDNA

2.0 at http://sparks.informatics.iupui.edu. The server version was trained by the

entire 212 protein-DNA complexes.

2009 Wiley-Liss, Inc.

DOI: 10.1002/prot.22384

PMCID: PMC2743280

PMID: 19274740 [Indexed for MEDLINE]

2153. Bioinformatics. 2009 Aug 1;25(15):1972-3. doi: 10.1093/bioinformatics/btp348.

Epub 2009 Jun 8.

trimAl: a tool for automated alignment trimming in large-scale phylogenetic

analyses.

Capella-Gutiérrez S(1), Silla-Martínez JM, Gabaldón T.

Author information:

(1)Comparative Genomics Group, Bioinformatics and Genomics Programme, Centre for

Genomic Regulation, 88 08003 Barcelona, Spain.

SUMMARY: Multiple sequence alignments are central to many areas of

bioinformatics. It has been shown that the removal of poorly aligned regions from

an alignment increases the quality of subsequent analyses. Such an alignment

trimming phase is complicated in large-scale phylogenetic analyses that deal with

thousands of alignments. Here, we present trimAl, a tool for automated alignment

trimming, which is especially suited for large-scale phylogenetic analyses.

trimAl can consider several parameters, alone or in multiple combinations, for

selecting the most reliable positions in the alignment. These include the

proportion of sequences with a gap, the level of amino acid similarity and, if

several alignments for the same set of sequences are provided, the level of

consistency across different alignments. Moreover, trimAl can automatically

select the parameters to be used in each specific alignment so that the

signal-to-noise ratio is optimized.

AVAILABILITY: trimAl has been written in C++, it is portable to all platforms.

trimAl is freely available for download (http://trimal.cgenomics.org) and can be

used online through the Phylemon web server (http://phylemon2.bioinfo.cipf.es/).

Supplementary Material is available at http://trimal.cgenomics.org/publications.

DOI: 10.1093/bioinformatics/btp348

PMCID: PMC2712344

PMID: 19505945 [Indexed for MEDLINE]

2154. Bioinformatics. 2009 Aug 1;25(15):1987-8. doi: 10.1093/bioinformatics/btp268.

Epub 2009 May 6.

CycSim--an online tool for exploring and experimenting with genome-scale

metabolic models.

Le Fèvre F(1), Smidtas S, Combe C, Durot M, d'Alché-Buc F, Schachter V.

Author information:

(1)CEA, DSV, IG, Genoscope, UMR 8030, Evry F-91057, France.

SUMMARY: CycSim is a web application dedicated to in silico experiments with

genome-scale metabolic models coupled to the exploration of knowledge from BioCyc

and KEGG. Specifically, CycSim supports the design of knockout experiments:

simulation of growth phenotypes of single or multiple gene deletions mutants on

specified media, comparison of these predictions with experimental phenotypes and

direct visualization of both on metabolic maps. The web interface is designed for

simplicity, putting constraint-based modelling techniques within easier reach of

biologists. CycSim also functions as an online repository of genome-scale

metabolic models.

AVAILABILITY: http://www.genoscope.cns.fr/cycsim.

DOI: 10.1093/bioinformatics/btp268

PMCID: PMC2712333

PMID: 19420054 [Indexed for MEDLINE]

2155. Bioinformatics. 2009 Aug 1;25(15):1989-90. doi: 10.1093/bioinformatics/btp287.

Epub 2009 May 4.

MetNetAligner: a web service tool for metabolic network alignments.

Cheng Q(1), Harrison R, Zelikovsky A.

Author information:

(1)Department of Computer Science, Georgia State University, Atlanta, GA 30303,

USA.

SUMMARY: The accumulation of high-throughput genomic, proteomic and metabolical

data allows for increasingly accurate modeling and reconstruction of metabolic

networks. Alignment of the reconstructed networks can help to catch model

inconsistencies and infer missing elements. In this note, we present the web

service tool MetNetAligner which aligns metabolic networks, taking in account the

similarity of network topology and the enzymes' functions. It can be used for

predicting unknown pathways, comparing and finding conserved patterns and

resolving ambiguous identification of enzymes. The tool supports several

alignment options including allowing or forbidding enzyme deletion and insertion.

It is based on a novel scoring scheme which measures enzyme-to-enzyme functional

similarity and a fast algorithm which efficiently finds optimal mappings from a

directed graph with restricted cyclic structure to an arbitrary directed graph.

AVAILABILITY: MetNetAligner is available as web-server at:

http://alla.cs.gsu.edu:8080/MinePW/pages/gmapping/GMMain.html.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp287

PMID: 19414533 [Indexed for MEDLINE]

2156. Bioinformatics. 2009 Aug 1;25(15):1974-5. doi: 10.1093/bioinformatics/btp250.

Epub 2009 Apr 27.

VARNA: Interactive drawing and editing of the RNA secondary structure.

Darty K(1), Denise A, Ponty Y.

Author information:

(1)LRI, UMR CNRS 8623, UMR CNRS 8621, Université Paris-Sud 11, F91405 Orsay

cedex, France. ponty@lri.fr

DESCRIPTION: VARNA is a tool for the automated drawing, visualization and

annotation of the secondary structure of RNA, designed as a companion software

for web servers and databases.

FEATURES: VARNA implements four drawing algorithms, supports input/output using

the classic formats dbn, ct, bpseq and RNAML and exports the drawing as five

picture formats, either pixel-based (JPEG, PNG) or vector-based (SVG, EPS and

XFIG). It also allows manual modification and structural annotation of the

resulting drawing using either an interactive point and click approach, within a

web server or through command-line arguments.

AVAILABILITY: VARNA is a free software, released under the terms of the GPLv3.0

license and available at http://varna.lri.fr.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp250

PMCID: PMC2712331

PMID: 19398448 [Indexed for MEDLINE]

2157. Comput Biol Chem. 2009 Aug;33(4):245-52. doi: 10.1016/j.compbiolchem.2009.04.006.

Epub 2009 May 9.

Computation of direct and inverse mutations with the SEGM web server (Stochastic

Evolution of Genetic Motifs): an application to splice sites of human genome

introns.

Benard E(1), Michel CJ.

Author information:

(1)Université de Strasbourg, LSIIT (UMR ULP-CNRS 7005), FDBT, Equipe de

Bioinformatique Théorique, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch,

France.

We present here the SEGM web server (Stochastic Evolution of Genetic Motifs) in

order to study the evolution of genetic motifs both in the direct evolutionary

sense (past-present) and in the inverse evolutionary sense (present-past). The

genetic motifs studied can be nucleotides, dinucleotides and trinucleotides. As

an example of an application of SEGM and to understand its functionalities, we

give an analysis of inverse mutations of splice sites of human genome introns.

SEGM is freely accessible at

http://lsiit-bioinfo.u-strasbg.fr:8080/webMathematica/SEGM/SEGM.html directly or

by the web site http://dpt-info.u-strasbg.fr/~michel/. To our knowledge, this

SEGM web server is to date the only computational biology software in this

evolutionary approach.

DOI: 10.1016/j.compbiolchem.2009.04.006

PMID: 19535295 [Indexed for MEDLINE]

2158. Nucleic Acids Res. 2009 Aug;37(15):e101. doi: 10.1093/nar/gkp491. Epub 2009 Jun

9.

QUBIC: a qualitative biclustering algorithm for analyses of gene expression data.

Li G(1), Ma Q, Tang H, Paterson AH, Xu Y.

Author information:

(1)Department of Biochemistry and Molecular Biology, Computational Systems

Biology Laboratory, Institute of Bioinformatics, University of Georgia, Athens,

GA 30602, USA.

Biclustering extends the traditional clustering techniques by attempting to find

(all) subgroups of genes with similar expression patterns under to-be-identified

subsets of experimental conditions when applied to gene expression data. Still

the real power of this clustering strategy is yet to be fully realized due to the

lack of effective and efficient algorithms for reliably solving the general

biclustering problem. We report a QUalitative BIClustering algorithm (QUBIC) that

can solve the biclustering problem in a more general form, compared to existing

algorithms, through employing a combination of qualitative (or semi-quantitative)

measures of gene expression data and a combinatorial optimization technique. One

key unique feature of the QUBIC algorithm is that it can identify all

statistically significant biclusters including biclusters with the so-called

'scaling patterns', a problem considered to be rather challenging; another key

unique feature is that the algorithm solves such general biclustering problems

very efficiently, capable of solving biclustering problems with tens of thousands

of genes under up to thousands of conditions in a few minutes of the CPU time on

a desktop computer. We have demonstrated a considerably improved biclustering

performance by our algorithm compared to the existing algorithms on various

benchmark sets and data sets of our own. QUBIC was written in ANSI C and tested

using GCC (version 4.1.2) on Linux. Its source code is available at:

http://csbl.bmb.uga.edu/ approximately maqin/bicluster. A server version of QUBIC

is also available upon request.

DOI: 10.1093/nar/gkp491

PMCID: PMC2731891

PMID: 19509312 [Indexed for MEDLINE]

2159. Proteins. 2009 Aug 1;76(2):343-52. doi: 10.1002/prot.22349.

Accurate domain identification with structure-anchored hidden Markov models,

saHMMs.

Tångrot JE(1), Kågström B, Sauer UH.

Author information:

(1)Umeå Centre for Molecular Pathogenesis, UCMP, Umeå University, Sweden.

The ever increasing speed of DNA sequencing widens the discrepancy between the

number of known gene products, and the knowledge of their function and structure.

Proper annotation of protein sequences is therefore crucial if the missing

information is to be deduced from sequence-based similarity comparisons. These

comparisons become exceedingly difficult as the pairwise identities drop to very

low values. To improve the accuracy of domain identification, we exploit the fact

that the three-dimensional structures of domains are much more conserved than

their sequences. Based on structure-anchored multiple sequence alignments of low

identity homologues we constructed 850 structure-anchored hidden Markov models

(saHMMs), each representing one domain family. Since the saHMMs are highly family

specific, they can be used to assign a domain to its correct family and clearly

distinguish it from domains belonging to other families, even within the same

superfamily. This task is not trivial and becomes particularly difficult if the

unknown domain is distantly related to the rest of the domain sequences within

the family. In a search with full length protein sequences, harbouring at least

one domain as defined by the structural classification of proteins database

(SCOP), version 1.71, versus the saHMM database based on SCOP version 1.69, we

achieve an accuracy of 99.0%. All of the few hits outside the family fall within

the correct superfamily. Compared to Pfam\_ls HMMs, the saHMMs obtain about 11%

higher coverage. A comparison with BLAST and PSI-BLAST demonstrates that the

saHMMs have consistently fewer errors per query at a given coverage. Within our

recommended E-value range, the same is true for a comparison with SUPERFAMILY.

Furthermore, we are able to annotate 232 proteins with 530 nonoverlapping domains

belonging to 102 different domain families among human proteins labelled

"unknown" in the NCBI protein database. Our results demonstrate that the saHMM

database represents a versatile and reliable tool for identification of domains

in protein sequences. With the aid of saHMMs, homology on the family level can be

assigned, even for distantly related sequences. Due to the construction of the

saHMMs, the hits they provide are always associated with high quality crystal

structures. The saHMM database can be accessed via the FISH server at

http://babel.ucmp.umu.se/fish/.

2008 Wiley-Liss, Inc.

DOI: 10.1002/prot.22349

PMID: 19173309 [Indexed for MEDLINE]

2160. BMC Med Inform Decis Mak. 2009 Jul 29;9:36. doi: 10.1186/1472-6947-9-36.

FluDetWeb: an interactive web-based system for the early detection of the onset

of influenza epidemics.

Conesa D(1), López-Quílez A, Martínez-Beneito MA, Miralles MT, Verdejo F.

Author information:

(1)Departament d'Estadística i Investigació Operativa, Universitat de València,

46100 Burjassot, Valencia, Spain. david.v.conesa@uv.es

BACKGROUND: The early identification of influenza outbreaks has became a priority

in public health practice. A large variety of statistical algorithms for the

automated monitoring of influenza surveillance have been proposed, but most of

them require not only a lot of computational effort but also operation of

sometimes not-so-friendly software.

RESULTS: In this paper, we introduce FluDetWeb, an implementation of a

prospective influenza surveillance methodology based on a client-server

architecture with a thin (web-based) client application design. Users can

introduce and edit their own data consisting of a series of weekly influenza

incidence rates. The system returns the probability of being in an epidemic phase

(via e-mail if desired). When the probability is greater than 0.5, it also

returns the probability of an increase in the incidence rate during the following

week. The system also provides two complementary graphs. This system has been

implemented using statistical free-software (R and WinBUGS), a web server

environment for Java code (Tomcat) and a software module created by us (Rdp)

responsible for managing internal tasks; the software package MySQL has been used

to construct the database management system. The implementation is available

on-line from: http://www.geeitema.org/meviepi/fludetweb/.

CONCLUSION: The ease of use of FluDetWeb and its on-line availability can make it

a valuable tool for public health practitioners who want to obtain information

about the probability that their system is in an epidemic phase. Moreover, the

architecture described can also be useful for developers of systems based on

computationally intensive methods.

DOI: 10.1186/1472-6947-9-36

PMCID: PMC2732617

PMID: 19640304 [Indexed for MEDLINE]

2161. Bioinformation. 2009 Jul 27;3(10):413-4.

DECOMP: a PDB decomposition tool on the web.

Ordog R(1), Szabadka Z, Grolmusz V.

Author information:

(1)Protein Information Technology Group, Eötvös University, 1117 Budapest,

Hungary.

The protein databank (PDB) contains high quality structural data for

computational structural biology investigations. We have earlier described a fast

tool (the decomp\_pdb tool) for identifying and marking missing atoms and residues

in PDB files. The tool also automatically decomposes PDB entries into separate

files describing ligands and polypeptide chains. Here, we describe a web

interface named DECOMP for the tool. Our program correctly identifies

multi-monomer ligands, and the server also offers the preprocessed ligand-protein

decomposition of the complete PDB for downloading (up to size: 5GB) AVAILABILITY:

http://decomp.pitgroup.org.

PMCID: PMC2737496

PMID: 19759860

2162. Front Neuroinform. 2009 Jul 20;3:22. doi: 10.3389/neuro.11.022.2009. eCollection

2009.

Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI

Pipeline.

Dinov ID(1), Van Horn JD, Lozev KM, Magsipoc R, Petrosyan P, Liu Z,

Mackenzie-Graham A, Eggert P, Parker DS, Toga AW.

Author information:

(1)Laboratory of Neuro Imaging, University of California Los Angeles, CA, USA.

The LONI Pipeline is a graphical environment for construction, validation and

execution of advanced neuroimaging data analysis protocols (Rex et al., 2003). It

enables automated data format conversion, allows Grid utilization, facilitates

data provenance, and provides a significant library of computational tools. There

are two main advantages of the LONI Pipeline over other graphical analysis

workflow architectures. It is built as a distributed Grid computing environment

and permits efficient tool integration, protocol validation and broad resource

distribution. To integrate existing data and computational tools within the LONI

Pipeline environment, no modification of the resources themselves is required.

The LONI Pipeline provides several types of process submissions based on the

underlying server hardware infrastructure. Only workflow instructions and

references to data, executable scripts and binary instructions are stored within

the LONI Pipeline environment. This makes it portable, computationally efficient,

distributed and independent of the individual binary processes involved in

pipeline data-analysis workflows. We have expanded the LONI Pipeline (V.4.2) to

include server-to-server (peer-to-peer) communication and a 3-tier failover

infrastructure (Grid hardware, Sun Grid Engine/Distributed Resource Management

Application API middleware, and the Pipeline server). Additionally, the LONI

Pipeline provides three layers of background-server executions for all

users/sites/systems. These new LONI Pipeline features facilitate

resource-interoperability, decentralized computing, construction and validation

of efficient and robust neuroimaging data-analysis workflows. Using brain imaging

data from the Alzheimer's Disease Neuroimaging Initiative (Mueller et al., 2005),

we demonstrate integration of disparate resources, graphical construction of

complex neuroimaging analysis protocols and distributed parallel computing. The

LONI Pipeline, its features, specifications, documentation and usage are

available online (http://Pipeline.loni.ucla.edu).

DOI: 10.3389/neuro.11.022.2009

PMCID: PMC2718780

PMID: 19649168

2163. BMC Bioinformatics. 2009 Jul 17;10:222. doi: 10.1186/1471-2105-10-222.

PCI-SS: MISO dynamic nonlinear protein secondary structure prediction.

Green JR(1), Korenberg MJ, Aboul-Magd MO.

Author information:

(1)Department of Systems and Computer Engineering, Carleton University, Ottawa,

Ontario, Canada. jrgreen@sce.carleton.ca

BACKGROUND: Since the function of a protein is largely dictated by its three

dimensional configuration, determining a protein's structure is of fundamental

importance to biology. Here we report on a novel approach to determining the one

dimensional secondary structure of proteins (distinguishing alpha-helices,

beta-strands, and non-regular structures) from primary sequence data which makes

use of Parallel Cascade Identification (PCI), a powerful technique from the field

of nonlinear system identification.

RESULTS: Using PSI-BLAST divergent evolutionary profiles as input data, dynamic

nonlinear systems are built through a black-box approach to model the process of

protein folding. Genetic algorithms (GAs) are applied in order to optimize the

architectural parameters of the PCI models. The three-state prediction problem is

broken down into a combination of three binary sub-problems and protein structure

classifiers are built using 2 layers of PCI classifiers. Careful construction of

the optimization, training, and test datasets ensures that no homology exists

between any training and testing data. A detailed comparison between PCI and 9

contemporary methods is provided over a set of 125 new protein chains guaranteed

to be dissimilar to all training data. Unlike other secondary structure

prediction methods, here a web service is developed to provide both human- and

machine-readable interfaces to PCI-based protein secondary structure prediction.

This server, called PCI-SS, is available at http://bioinf.sce.carleton.ca/PCISS.

In addition to a dynamic PHP-generated web interface for humans, a Simple Object

Access Protocol (SOAP) interface is added to permit invocation of the PCI-SS

service remotely. This machine-readable interface facilitates incorporation of

PCI-SS into multi-faceted systems biology analysis pipelines requiring protein

secondary structure information, and greatly simplifies high-throughput analyses.

XML is used to represent the input protein sequence data and also to encode the

resulting structure prediction in a machine-readable format. To our knowledge,

this represents the only publicly available SOAP-interface for a protein

secondary structure prediction service with published WSDL interface definition.

CONCLUSION: Relative to the 9 contemporary methods included in the comparison

cascaded PCI classifiers perform well, however PCI finds greatest application as

a consensus classifier. When PCI is used to combine a sequence-to-structure

PCI-based classifier with the current leading ANN-based method, PSIPRED, the

overall error rate (Q3) is maintained while the rate of occurrence of a

particularly detrimental error is reduced by up to 25%. This improvement in BAD

score, combined with the machine-readable SOAP web service interface makes PCI-SS

particularly useful for inclusion in a tertiary structure prediction pipeline.

DOI: 10.1186/1471-2105-10-222

PMCID: PMC2720391

PMID: 19615046 [Indexed for MEDLINE]

2164. Bioinformatics. 2009 Jul 15;25(14):1739-45. doi: 10.1093/bioinformatics/btp309.

Epub 2009 May 12.

ESG: extended similarity group method for automated protein function prediction.

Chitale M(1), Hawkins T, Park C, Kihara D.

Author information:

(1)Department of Computer Science, Purdue University, IN 47907, USA.

MOTIVATION: Importance of accurate automatic protein function prediction is ever

increasing in the face of a large number of newly sequenced genomes and

proteomics data that are awaiting biological interpretation. Conventional methods

have focused on high sequence similarity-based annotation transfer which relies

on the concept of homology. However, many cases have been reported that simple

transfer of function from top hits of a homology search causes erroneous

annotation. New methods are required to handle the sequence similarity in a more

robust way to combine together signals from strongly and weakly similar proteins

for effectively predicting function for unknown proteins with high reliability.

RESULTS: We present the extended similarity group (ESG) method, which performs

iterative sequence database searches and annotates a query sequence with Gene

Ontology terms. Each annotation is assigned with probability based on its

relative similarity score with the multiple-level neighbors in the protein

similarity graph. We will depict how the statistical framework of ESG improves

the prediction accuracy by iteratively taking into account the neighborhood of

query protein in the sequence similarity space. ESG outperforms conventional

PSI-BLAST and the protein function prediction (PFP) algorithm. It is found that

the iterative search is effective in capturing multiple-domains in a query

protein, enabling accurately predicting several functions which originate from

different domains.

AVAILABILITY: ESG web server is available for automated protein function

prediction at http://dragon.bio.purdue.edu/ESG/.

DOI: 10.1093/bioinformatics/btp309

PMCID: PMC2705228

PMID: 19435743 [Indexed for MEDLINE]

2165. Bioinformatics. 2009 Jul 15;25(14):1846-8. doi: 10.1093/bioinformatics/btp293.

Epub 2009 May 4.

PESTAS: a web server for EST analysis and sequence mining.

Nam SH(1), Kim DW, Jung TS, Choi YS, Kim DW, Choi HS, Choi SH, Park HS.

Author information:

(1)Industrial Biotechnology & Bioenergy Research Center, Korea Research Institute

of Bioscience and Biotechnology, Daejeon 305-806, Korea.

SUMMARY: We have developed a web server for the high-throughput annotation of

expressed sequence tags (ESTs) called pipeline for EST analysis service (PESTAS).

PESTAS processes entire datasets with an automated pipeline of 13 analytic

services, then deposits the data into the MySQL database and transforms it into

three kinds of reports: preprocessing, assembling and annotation. All annotated

information is provided to the scientist and can be downloaded through a web

browser. To get more relevant functional annotation results, a curation function

was introduced with which biologists can easily change the best-hit annotation

information. We included a gene chip module that detects gene expression

differences between libraries by comparing accession number counts from BLAST

search results. PESTAS also provides access to the pathway information of KEGG,

which is useful for mapping the relationships among networks of annotated

enzymes, and is especially valuable for those researchers interested in

biological pathways.

AVAILABILITY: PESTAS is available at http://pestas.kribb.re.kr/.

DOI: 10.1093/bioinformatics/btp293

PMID: 19414531 [Indexed for MEDLINE]

2166. J Comput Chem. 2009 Jul 15;30(9):1532-43. doi: 10.1002/jcc.21232.

Incorporating structural characteristics for identification of protein

methylation sites.

Shien DM(1), Lee TY, Chang WC, Hsu JB, Horng JT, Hsu PC, Wang TY, Huang HD.

Author information:

(1)Department of Computer Science and Information Engineering, National Central

University, Chung-Li 320, Taiwan.

Studies over the last few years have identified protein methylation on histones

and other proteins that are involved in the regulation of gene transcription.

Several works have developed approaches to identify computationally the potential

methylation sites on lysine and arginine. Studies of protein tertiary structure

have demonstrated that the sites of protein methylation are preferentially in

regions that are easily accessible. However, previous studies have not taken into

account the solvent-accessible surface area (ASA) that surrounds the methylation

sites. This work presents a method named MASA that combines the support vector

machine with the sequence and structural characteristics of proteins to identify

methylation sites on lysine, arginine, glutamate, and asparagine. Since most

experimental methylation sites are not associated with corresponding protein

tertiary structures in the Protein Data Bank, the effective solvent-accessible

prediction tools have been adopted to determine the potential ASA values of amino

acids in proteins. Evaluation of predictive performance by cross-validation

indicates that the ASA values around the methylation sites can improve the

accuracy of prediction. Additionally, an independent test reveals that the

prediction accuracies for methylated lysine and arginine are 80.8 and 85.0%,

respectively. Finally, the proposed method is implemented as an effective system

for identifying protein methylation sites. The developed web server is freely

available at http://MASA.mbc.nctu.edu.tw/.

(c) 2009 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.21232

PMID: 19263424 [Indexed for MEDLINE]

2167. J Comput Chem. 2009 Jul 15;30(9):1414-23. doi: 10.1002/jcc.21163.

GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled

receptor functional classes.

Xiao X(1), Wang P, Chou KC.

Author information:

(1)Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 33300,

China. xiaoxuan0326@yahoo.com.cn

Given an uncharacterized protein sequence, how can we identify whether it is a

G-protein-coupled receptor (GPCR) or not? If it is, which functional family class

does it belong to? It is important to address these questions because GPCRs are

among the most frequent targets of therapeutic drugs and the information thus

obtained is very useful for "comparative and evolutionary pharmacology," a

technique often used for drug development. Here, we present a web-server

predictor called "GPCR-CA," where "CA" stands for "Cellular Automaton" (Wolfram,

S. Nature 1984, 311, 419), meaning that the CA images have been utilized to

reveal the pattern features hidden in piles of long and complicated protein

sequences. Meanwhile, the gray-level co-occurrence matrix factors extracted from

the CA images are used to represent the samples of proteins through their pseudo

amino acid composition (Chou, K.C. Proteins 2001, 43, 246). GPCR-CA is a

two-layer predictor: the first layer prediction engine is for identifying a query

protein as GPCR on non-GPCR; if it is a GPCR protein, the process will be

automatically continued with the second-layer prediction engine to further

identify its type among the following six functional classes: (a) rhodopsin-like,

(b) secretin-like, (c) metabotrophic/glutamate/pheromone; (d) fungal pheromone,

(e) cAMP receptor, and (f) frizzled/smoothened family. The overall success rates

by the predictor for the first and second layers are over 91% and 83%,

respectively, that were obtained through rigorous jackknife cross-validation

tests on a new-constructed stringent benchmark dataset in which none of proteins

has >or=40% pairwise sequence identity to any other in a same subset. GPCR-CA is

freely accessible at http://218.65.61.89:8080/bioinfo/GPCR-CA, by which one can

get the desired two-layer results for a query protein sequence within about 20

seconds.

(c) 2008 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.21163

PMID: 19037861 [Indexed for MEDLINE]

2168. PLoS One. 2009 Jul 15;4(7):e6176. doi: 10.1371/journal.pone.0006176.

ProteinArchitect: protein evolution above the sequence level.

Haimel M(1), Pröll K, Rebhan M.

Author information:

(1)Department of Bioinformatics, Upper Austrian University of Applied Sciences,

Hagenberg, Austria.

BACKGROUND: While many authors have discussed models and tools for studying

protein evolution at the sequence level, molecular function is usually mediated

by complex, higher order features such as independently folding domains and

linear motifs that are based on or embedded in a particular arrangment of

features such as secondary structure elements, transmembrane domains and regions

with intrinsic disorder. This 'protein architecture' can, in its most simplistic

representation, be visualized as domain organization cartoons that can be used to

compare proteins in terms of the order of their mostly globular domains.

METHODOLOGY: Here, we describe a visual approach and a webserver for protein

comparison that extend the domain organization cartoon concept. By developing an

information-rich, compact visualization of different protein features above the

sequence level, potentially related proteins can be compared at the level of

propensities for secondary structure, transmembrane domains and intrinsic

disorder, in addition to PFAM domains. A public Web server is available at

www.proteinarchitect.net, while the code is provided at

protarchitect.sourceforge.net.

CONCLUSIONS/SIGNIFICANCE: Due to recent advances in sequencing technologies we

are now flooded with millions of predicted proteins that await comparative

analysis. In many cases, mature tools focused on revealing hits with considerable

global or local similarity to well-characterized proteins will not be able to

lead us to testable hypotheses about a protein's function, or the function of a

particular region. The visual comparison of different types of protein features

with ProteinArchitect will be useful when assessing the relevance of similarity

search hits, to discover subgroups in protein families and superfamilies, and to

understand protein regions with conserved features outside globular regions.

Therefore, this approach is likely to help researchers to develop testable

hypotheses about a protein's function even if is somewhat distant from the more

characterized proteins, by facilitating the discovery of features that are

conserved above the sequence level for comparison and further experimental

investigation.

DOI: 10.1371/journal.pone.0006176

PMCID: PMC2705671

PMID: 19603068 [Indexed for MEDLINE]

2169. BMC Genomics. 2009 Jul 14;10 Suppl 2:S6. doi: 10.1186/1471-2164-10-S2-S6.

BioNetBuilder2.0: bringing systems biology to chicken and other model organisms.

Konieczka JH(1), Drew K, Pine A, Belasco K, Davey S, Yatskievych TA, Bonneau R,

Antin PB.

Author information:

(1)Department of Molecular & Cellular Biology, University of Arizona, Tucson, AZ,

USA. jayk@u.arizona.edu.

BACKGROUND: Systems Biology research tools, such as Cytoscape, have greatly

extended the reach of genomic research. By providing platforms to integrate data

with molecular interaction networks, researchers can more rapidly begin

interpretation of large data sets collected for a system of interest.

BioNetBuilder is an open-source client-server Cytoscape plugin that automatically

integrates molecular interactions from all major public interaction databases and

serves them directly to the user's Cytoscape environment. Until recently however,

chicken and other eukaryotic model systems had little interaction data available.

RESULTS: Version 2.0 of BioNetBuilder includes a redesigned synonyms resolution

engine that enables transfer and integration of interactions across species; this

engine translates between alternate gene names as well as between orthologs in

multiple species. Additionally, BioNetBuilder is now implemented to be part of

the Gaggle, thereby allowing seamless communication of interaction data to any

software implementing the widely used Gaggle software. Using BioNetBuilder, we

constructed a chicken interactome possessing 72,000 interactions among 8,140

genes directly in the Cytoscape environment. In this paper, we present a tutorial

on how to do so and analysis of a specific use case.

CONCLUSION: BioNetBuilder 2.0 provides numerous user-friendly systems biology

tools that were otherwise inaccessible to researchers in chicken genomics, as

well as other model systems. We provide a detailed tutorial spanning all required

steps in the analysis. BioNetBuilder 2.0, the tools for maintaining its data

bases, standard operating procedures for creating local copies of its back-end

data bases, as well as all of the Gaggle and Cytoscape codes required, are

open-source and freely available at

http://err.bio.nyu.edu/cytoscape/bionetbuilder/.

DOI: 10.1186/1471-2164-10-S2-S6

PMCID: PMC2966329

PMID: 19607657 [Indexed for MEDLINE]

2170. BMC Struct Biol. 2009 Jul 9;9:44. doi: 10.1186/1472-6807-9-44.

Amyloidogenic determinants are usually not buried.

Frousios KK(1), Iconomidou VA, Karletidi CM, Hamodrakas SJ.

Author information:

(1)Department of Cell Biology and Biophysics, Faculty of Biology, University of

Athens, Panepistimiopolis, Athens 15701, Greece. kfrousios@biol.uoa.gr

BACKGROUND: Amyloidoses are a group of usually fatal diseases, probably caused by

protein misfolding and subsequent aggregation into amyloid fibrillar deposits.

The mechanisms involved in amyloid fibril formation are largely unknown and are

the subject of current, intensive research. In an attempt to identify possible

amyloidogenic regions in proteins for further experimental investigation, we have

developed and present here a publicly available online tool that utilizes five

different and independently published methods, to form a consensus prediction of

amyloidogenic regions in proteins, using only protein primary structure data.

RESULTS: It appears that the consensus prediction tool is slightly more objective

than individual prediction methods alone and suggests several previously not

identified amino acid stretches as potential amyloidogenic determinants, which

(although several of them may be overpredictions) require further experimental

studies. The tool is available at: http://biophysics.biol.uoa.gr/AMYLPRED.

Utilizing molecular graphics programs, like O and PyMOL, as well as the algorithm

DSSP, it was found that nearly all experimentally verified amyloidogenic

determinants (short peptide stretches favouring aggregation and subsequent

amyloid formation), and several predicted, with the aid of the tool AMYLPRED, but

not experimentally verified amyloidogenic determinants, are located on the

surface of the relevant amyloidogenic proteins. This finding may be important in

efforts directed towards inhibiting amyloid fibril formation.

CONCLUSION: The most significant result of this work is the observation that

virtually all, to date, experimentally determined amyloidogenic determinants and

the majority of predicted, but not yet experimentally verified short

amyloidogenic stretches, lie 'exposed' on the surface of the relevant

amyloidogenic proteins, and also several of them have the ability to act as

conformational 'switches'. Experiments, focused on these fragments, should be

performed to test this idea.

DOI: 10.1186/1472-6807-9-44

PMCID: PMC2714319

PMID: 19589171 [Indexed for MEDLINE]

2171. BMC Genomics. 2009 Jul 7;10 Suppl 1:S1. doi: 10.1186/1471-2164-10-S1-S1.

Prediction of DNA-binding residues from protein sequence information using random

forests.

Wang L(1), Yang MQ, Yang JY.

Author information:

(1)Department of Genetics and Biochemistry, Clemson University, Clemson, SC

29634, USA. liangjw@clemson.edu

BACKGROUND: Protein-DNA interactions are involved in many biological processes

essential for cellular function. To understand the molecular mechanism of

protein-DNA recognition, it is necessary to identify the DNA-binding residues in

DNA-binding proteins. However, structural data are available for only a few

hundreds of protein-DNA complexes. With the rapid accumulation of sequence data,

it becomes an important but challenging task to accurately predict DNA-binding

residues directly from amino acid sequence data.

RESULTS: A new machine learning approach has been developed in this study for

predicting DNA-binding residues from amino acid sequence data. The approach used

both the labelled data instances collected from the available structures of

protein-DNA complexes and the abundant unlabeled data found in protein sequence

databases. The evolutionary information contained in the unlabeled sequence data

was represented as position-specific scoring matrices (PSSMs) and several new

descriptors. The sequence-derived features were then used to train random forests

(RFs), which could handle a large number of input variables and avoid model

overfitting. The use of evolutionary information was found to significantly

improve classifier performance. The RF classifier was further evaluated using a

separate test dataset, and the predicted DNA-binding residues were examined in

the context of three-dimensional structures.

CONCLUSION: The results suggest that the RF-based approach gives rise to more

accurate prediction of DNA-binding residues than previous studies. A new web

server called BindN-RF http://bioinfo.ggc.org/bindn-rf/ has thus been developed

to make the RF classifier accessible to the biological research community.

DOI: 10.1186/1471-2164-10-S1-S1

PMCID: PMC2709252

PMID: 19594868 [Indexed for MEDLINE]

2172. Bioinform Biol Insights. 2009 Jul 1;3:71-81.

Automated detection of conformational epitopes using phage display Peptide

sequences.

Negi SS(1), Braun W.

Author information:

(1)Sealy Center for Structural Biology and Molecular Biophysics, Department of

Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston

TX, 77555-0857, USA. Email: webraun@utmb.edu.

BACKGROUND: Precise determination of conformational epitopes of neutralizing

antibodies represents a key step in the rational design of novel vaccines. A

powerful experimental method to gain insights on the physical chemical nature of

conformational epitopes is the selection of linear peptides that bind with high

affinities to a monoclonal antibody of interest by phage display technology.

However, the structural characterization of conformational epitopes from these

mimotopes is not straightforward, and in the past the interpretation of peptide

sequences from phage display experiments focused on linear sequence analysis to

find a consensus sequence or common sequence motifs.

RESULTS: We present a fully automated search method, EpiSearch that predicts the

possible location of conformational epitopes on the surface of an antigen. The

algorithm uses peptide sequences from phage display experiments as input, and

ranks all surface exposed patches according to the frequency distribution of

similar residues in the peptides and in the patch. We have tested the performance

of the EpiSearch algorithm for six experimental data sets of phage display

experiments, the human epidermal growth factor receptor-2 (HER-2/neu), the

antibody mAb Bo2C11 targeting the C(2) domain of FVIII, antibodies mAb 17b and

mAb b12 of the HIV envelope protein gp120, mAb 13b5 targeting HIV-1 capsid

protein and 80R of the SARS coronavirus spike protein. In all these examples the

conformational epitopes as determined by the X-ray crystal structures of the

antibody-antigen complexes, were found within the highest scoring patches of

EpiSearch, covering in most cases more than 50% residues of experimental observed

conformational epitopes. Input options of the program include mapping of a single

peptide or a set of peptides on the antigen structure, and the results of the

calculation can be visualized on our interactive web server.

AVAILABILITY: Users can access the EpiSearch from our web server

http://curie.utmb.edu/episearch.html.

PMCID: PMC2808184

PMID: 20140073

2173. Bioinformatics. 2009 Jul 1;25(13):1709-10. doi: 10.1093/bioinformatics/btp304.

Epub 2009 May 7.

FlexServ: an integrated tool for the analysis of protein flexibility.

Camps J(1), Carrillo O, Emperador A, Orellana L, Hospital A, Rueda M, Cicin-Sain

D, D'Abramo M, Gelpí JL, Orozco M.

Author information:

(1)Joint IRB-BSC Program on Computational Biology, Institute of Research in

Biomedicine, Barcelona Science Park, Josep Samitier 1-5, Barcelona, Spain.

SUMMARY: FlexServ is a web-based tool for the analysis of protein flexibility.

The server incorporates powerful protocols for the coarse-grained determination

of protein dynamics using different versions of Normal Mode Analysis (NMA),

Brownian dynamics (BD) and Discrete Dynamics (DMD). It can also analyze user

provided trajectories. The server allows a complete analysis of flexibility using

a large variety of metrics, including basic geometrical analysis, B-factors,

essential dynamics, stiffness analysis, collectivity measures, Lindemann's

indexes, residue correlation, chain-correlations, dynamic domain determination,

hinge point detections, etc. Data is presented through a web interface as plain

text, 2D and 3D graphics.

AVAILABILITY: http://mmb.pcb.ub.es/FlexServ; http://www.inab.org

DOI: 10.1093/bioinformatics/btp304

PMID: 19429600 [Indexed for MEDLINE]

2174. Bioinformatics. 2009 Jul 1;25(13):1625-31. doi: 10.1093/bioinformatics/btp296.

Epub 2009 May 5.

Flexible structural protein alignment by a sequence of local transformations.

Rocha J(1), Segura J, Wilson RC, Dasgupta S.

Author information:

(1)Department of Mathematics and Computer Science, University of the Balearic

Islands, Palma, Spain. jairo@uib.es

MOTIVATION: Throughout evolution, homologous proteins have common regions that

stay semi-rigid relative to each other and other parts that vary in a more

noticeable way. In order to compare the increasing number of structures in the

PDB, flexible geometrical alignments are needed, that are reliable and easy to

use.

RESULTS: We present a protein structure alignment method whose main feature is

the ability to consider different rigid transformations at different sites,

allowing for deformations beyond a global rigid transformation. The performance

of the method is comparable with that of the best ones from 10 aligners tested,

regarding both the quality of the alignments with respect to hand curated ones,

and the classification ability. An analysis of some structure pairs from the

literature that need to be matched in a flexible fashion are shown. The use of a

series of local transformations can be exported to other classifiers, and a

future golden protein similarity measure could benefit from it.

AVAILABILITY: A public server for the program is available at

http://dmi.uib.es/ProtDeform/.

SUPPLEMENTARY INFORMATION: All data used, results and examples are available at

http://dmi.uib.es/people/jairo/bio/ProtDeform.

DOI: 10.1093/bioinformatics/btp296

PMCID: PMC2940242

PMID: 19417057 [Indexed for MEDLINE]

2175. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W281-6. doi: 10.1093/nar/gkp477.

Epub 2009 Jun 16.

RNAmutants: a web server to explore the mutational landscape of RNA secondary

structures.

Waldispühl J(1), Devadas S, Berger B, Clote P.

Author information:

(1)Department of Mathematics, MIT, Cambridge, MA 02139, USA.

The history and mechanism of molecular evolution in DNA have been greatly

elucidated by contributions from genetics, probability theory and

bioinformatics--indeed, mathematical developments such as Kimura's neutral

theory, Kingman's coalescent theory and efficient software such as BLAST,

ClustalW, Phylip, etc., provide the foundation for modern population genetics. In

contrast to DNA, the function of most noncoding RNA depends on tertiary

structure, experimentally known to be largely determined by secondary structure,

for which dynamic programming can efficiently compute the minimum free energy

secondary structure. For this reason, understanding the effect of pointwise

mutations in RNA secondary structure could reveal fundamental properties of

structural RNA molecules and improve our understanding of molecular evolution of

RNA. The web server RNAmutants provides several efficient tools to compute the

ensemble of low-energy secondary structures for all k-mutants of a given RNA

sequence, where k is bounded by a user-specified upper bound. As we have

previously shown, these tools can be used to predict putative deleterious

mutations and to analyze regulatory sequences from the hepatitis C and human

immunodeficiency genomes. Web server is available at

http://bioinformatics.bc.edu/clotelab/RNAmutants/, and downloadable binaries at

http://rnamutants.csail.mit.edu/.

DOI: 10.1093/nar/gkp477

PMCID: PMC2703890

PMID: 19531740 [Indexed for MEDLINE]

2176. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W3-5. doi: 10.1093/nar/gkp531.

Epub 2009 Jun 15.

Evolution in bioinformatic resources: 2009 update on the Bioinformatics Links

Directory.

Brazas MD(1), Yamada JT, Ouellette BF.

Author information:

(1)Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto,

Ontario, Canada M5G 0A3.

All of the life science research web servers published in this and previous

issues of Nucleic Acids Research, together with other useful tools, databases and

resources for bioinformatics and molecular biology research are freely accessible

online through the Bioinformatics Links Directory,

http://bioinformatics.ca/links\_directory/. Entirely dependent on user feedback

and community input, the Bioinformatics Links Directory exemplifies an open

access research tool and resource. With 112 websites featured in the July 2009

Web Server Issue of Nucleic Acids Research, the 2009 update brings the total

number of servers listed in the Bioinformatics Links Directory close to an

impressive 1400 links. A complete list of all links listed in this Nucleic Acids

Research 2009 Web Server Issue can be accessed online at

http://bioinfomatics.ca/links\_directory/narweb2009/. The 2009 update of the

Bioinformatics Links Directory, which includes the Web Server list and summaries,

is also available online at the Nucleic Acids Research website,

http://nar.oxfordjournals.org/.

DOI: 10.1093/nar/gkp531

PMCID: PMC2703910

PMID: 19528072 [Indexed for MEDLINE]

2177. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W57-62. doi: 10.1093/nar/gkp404.

Epub 2009 Jun 15.

OmicBrowse: a Flash-based high-performance graphics interface for genomic

resources.

Matsushima A(1), Kobayashi N, Mochizuki Y, Ishii M, Kawaguchi S, Endo TA, Umetsu

R, Makita Y, Toyoda T.

Author information:

(1)Bioinformatics and Systems Engineering division (BASE), RIKEN (The Institute

of Physical and Chemical Research), 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa

230-0045, Japan.

OmicBrowse is a genome browser designed as a scalable system for maintaining

numerous genome annotation datasets. It is an open source tool capable of

regulating multiple user data access to each dataset to allow multiple users to

have their own integrative view of both their unpublished and published datasets,

so that the maintenance costs related to supplying each collaborator exclusively

with their own private data are significantly reduced. OmicBrowse supports DAS1

imports and exports of annotations to Internet site servers worldwide. We also

provide a data-download named OmicDownload server that interactively selects

datasets and filters the data on the selected datasets. Our OmicBrowse server has

been freely available at http://omicspace.riken.jp/ since its launch in 2003. The

OmicBrowse source code is downloadable from

http://sourceforge.net/projects/omicbrowse/.

DOI: 10.1093/nar/gkp404

PMCID: PMC2703975

PMID: 19528066 [Indexed for MEDLINE]

2178. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W122-8. doi: 10.1093/nar/gkp438.

Epub 2009 Jun 5.

DASMIweb: online integration, analysis and assessment of distributed protein

interaction data.

Blankenburg H(1), Ramírez F, Büch J, Albrecht M.

Author information:

(1)Max Planck Institute for Informatics, Campus E1.4, 66123 Saarbrücken, Germany.

hagen.blankenburg@mpi-inf.mpg.de

In recent years, we have witnessed a substantial increase of the amount of

available protein interaction data. However, most data are currently not readily

accessible to the biologist at a single site, but scattered over multiple online

repositories. Therefore, we have developed the DASMIweb server that affords the

integration, analysis and qualitative assessment of distributed sources of

interaction data in a dynamic fashion. Since DASMIweb allows for querying many

different resources of protein and domain interactions simultaneously, it serves

as an important starting point for interactome studies and assists the user in

finding publicly accessible interaction data with minimal effort. The pool of

queried resources is fully configurable and supports the inclusion of own

interaction data or confidence scores. In particular, DASMIweb integrates

confidence measures like functional similarity scores to assess individual

interactions. The retrieved results can be exported in different file formats

like MITAB or SIF. DASMIweb is freely available at http://www.dasmiweb.de.

DOI: 10.1093/nar/gkp438

PMCID: PMC2703953

PMID: 19502495 [Indexed for MEDLINE]

2179. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W106-8. doi: 10.1093/nar/gkp474.

Epub 2009 Jun 2.

TORQUE: topology-free querying of protein interaction networks.

Bruckner S(1), Hüffner F, Karp RM, Shamir R, Sharan R.

Author information:

(1)The Blavatnik School of Computer Science, Tel Aviv University, 69978 Tel Aviv,

Israel. bruckner@tau.ac.il

TORQUE is a tool for cross-species querying of protein-protein interaction

networks. It aims to answer the following question: given a set of proteins

constituting a known complex or a pathway in one species, can a similar complex

or pathway be found in the protein network of another species? To this end,

Torque seeks a matching set of proteins that are sequence similar to the query

proteins and span a connected region of the target network, while allowing for

both insertions and deletions. Unlike existing approaches, TORQUE does not

require knowledge of the interconnections among the query proteins. It can handle

large queries of up to 25 proteins. The Torque web server is freely available for

use at http://www.cs.tau.ac.il/~bnet/torque.html.

DOI: 10.1093/nar/gkp474

PMCID: PMC2703961

PMID: 19491310 [Indexed for MEDLINE]

2180. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W185-8. doi: 10.1093/nar/gkp434.

Epub 2009 Jun 2.

OrthoSelect: a web server for selecting orthologous gene alignments from EST

sequences.

Schreiber F(1), Wörheide G, Morgenstern B.

Author information:

(1)Institut für Mikrobiologie und Genetik, Abteilung für Bioinformatik,

Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077, Göttingen, Germany.

Fschrei@gwdg.de

In the absence of whole genome sequences for many organisms, the use of expressed

sequence tags (EST) offers an affordable approach for researchers conducting

phylogenetic analyses to gain insight about the evolutionary history of

organisms. Reliable alignments for phylogenomic analyses are based on orthologous

gene sequences from different taxa. So far, researchers have not sufficiently

tackled the problem of the completely automated construction of such datasets.

Existing software tools are either semi-automated, covering only part of the

necessary data processing, or implemented as a pipeline, requiring the

installation and configuration of a cascade of external tools, which may be

time-consuming and hard to manage. To simplify data set construction for

phylogenomic studies, we set up a web server that uses our recently developed

OrthoSelect approach. To the best of our knowledge, our web server is the first

web-based EST analysis pipeline that allows the detection of orthologous gene

sequences in EST libraries and outputs orthologous gene alignments. Additionally,

OrthoSelect provides the user with an extensive results section that lists and

visualizes all important results, such as annotations, data matrices for each

gene/taxon and orthologous gene alignments. The web server is available at

http://orthoselect.gobics.de.

DOI: 10.1093/nar/gkp434

PMCID: PMC2703962

PMID: 19491309 [Indexed for MEDLINE]

2181. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W260-5. doi: 10.1093/nar/gkp433.

Epub 2009 May 29.

SARA: a server for function annotation of RNA structures.

Capriotti E(1), Marti-Renom MA.

Author information:

(1)Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de

Investigación Príncipe Felipe, Valencia, Spain.

Recent interest in non-coding RNA transcripts has resulted in a rapid increase of

deposited RNA structures in the Protein Data Bank. However, a characterization

and functional classification of the RNA structure and function space have only

been partially addressed. Here, we introduce the SARA program for pair-wise

alignment of RNA structures as a web server for structure-based RNA function

assignment. The SARA server relies on the SARA program, which aligns two RNA

structures based on a unit-vector root-mean-square approach. The likely accuracy

of the SARA alignments is assessed by three different P-values estimating the

statistical significance of the sequence, secondary structure and tertiary

structure identity scores, respectively. Our benchmarks, which relied on a set of

419 RNA structures with known SCOR structural class, indicate that at a negative

logarithm of mean P-value higher or equal than 2.5, SARA can assign the correct

or a similar SCOR class to 81.4% and 95.3% of the benchmark set, respectively.

The SARA server is freely accessible via the World Wide Web at

http://sgu.bioinfo.cipf.es/services/SARA/.

DOI: 10.1093/nar/gkp433

PMCID: PMC2703911

PMID: 19483098 [Indexed for MEDLINE]

2182. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W492-7. doi: 10.1093/nar/gkp403.

Epub 2009 May 29.

SAM-T08, HMM-based protein structure prediction.

Karplus K(1).

Author information:

(1)Department of Biomolecular Engineering, Baskin School of Engineering,

University of California, Santa Cruz, CA 95064, USA. karplus@soe.ucsc.edu

The SAM-T08 web server is a protein structure prediction server that provides

several useful intermediate results in addition to the final predicted 3D

structure: three multiple sequence alignments of putative homologs using

different iterated search procedures, prediction of local structure features

including various backbone and burial properties, calibrated E-values for the

significance of template searches of PDB and residue-residue contact predictions.

The server has been validated as part of the CASP8 assessment of structure

prediction as having good performance across all classes of predictions. The

SAM-T08 server is available at

http://compbio.soe.ucsc.edu/SAM\_T08/T08-query.html.

DOI: 10.1093/nar/gkp403

PMCID: PMC2703928

PMID: 19483096 [Indexed for MEDLINE]

2183. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W587-92. doi:

10.1093/nar/gkp435. Epub 2009 May 29.

VisHiC--hierarchical functional enrichment analysis of microarray data.

Krushevskaya D(1), Peterson H, Reimand J, Kull M, Vilo J.

Author information:

(1)Institute of Computer Science, University of Tartu, Liivi 2, 50409 Tartu,

Estonian.

Measuring gene expression levels with microarrays is one of the key technologies

of modern genomics. Clustering of microarray data is an important application, as

genes with similar expression profiles may be regulated by common pathways and

involved in related functions. Gene Ontology (GO) analysis and visualization

allows researchers to study the biological context of discovered clusters and

characterize genes with previously unknown functions. We present VisHiC

(Visualization of Hierarchical Clustering), a web server for clustering and

compact visualization of gene expression data combined with automated function

enrichment analysis. The main output of the analysis is a dendrogram and visual

heatmap of the expression matrix that highlights biologically relevant clusters

based on enriched GO terms, pathways and regulatory motifs. Clusters with most

significant enrichments are contracted in the final visualization, while less

relevant parts are hidden altogether. Such a dense representation of microarray

data gives a quick global overview of thousands of transcripts in many conditions

and provides a good starting point for further analysis. VisHiC is freely

available at http://biit.cs.ut.ee/vishic.

DOI: 10.1093/nar/gkp435

PMCID: PMC2703939

PMID: 19483095 [Indexed for MEDLINE]

2184. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W63-7. doi: 10.1093/nar/gkp430.

Epub 2009 May 27.

AutoClass@IJM: a powerful tool for Bayesian classification of heterogeneous data

in biology.

Achcar F(1), Camadro JM, Mestivier D.

Author information:

(1)Modeling in Integrative Biology Group, Jacques Monod Institute, UMR7592 CNRS

and Univ Paris-Diderot, Bâtiment Buffon, 15 rue Hélène Brion, 75205 Paris Cedex

13, France.

Recently, several theoretical and applied studies have shown that unsupervised

Bayesian classification systems are of particular relevance for biological

studies. However, these systems have not yet fully reached the biological

community mainly because there are few freely available dedicated computer

programs, and Bayesian clustering algorithms are known to be time consuming,

which limits their usefulness when using personal computers. To overcome these

limitations, we developed AutoClass@IJM, a computational resource with a web

interface to AutoClass, a powerful unsupervised Bayesian classification system

developed by the Ames Research Center at N.A.S.A. AutoClass has many powerful

features with broad applications in biological sciences: (i) it determines the

number of classes automatically, (ii) it allows the user to mix discrete and real

valued data, (iii) it handles missing values. End users upload their data sets

through our web interface; computations are then queued in our cluster server.

When the clustering is completed, an URL to the results is sent back to the user

by e-mail. AutoClass@IJM is freely available at:

http://ytat2.ijm.univ-paris-diderot.fr/AutoclassAtIJM.html.

DOI: 10.1093/nar/gkp430

PMCID: PMC2703914

PMID: 19474346 [Indexed for MEDLINE]

2185. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W240-6. doi: 10.1093/nar/gkp358.

Epub 2009 May 27.

Web 3DNA--a web server for the analysis, reconstruction, and visualization of

three-dimensional nucleic-acid structures.

Zheng G(1), Lu XJ, Olson WK.

Author information:

(1)Department of Chemistry & Chemical Biology, BioMaPS Institute for Quantitative

Biology, Wright-Rieman Laboratories, Rutgers, State University of New Jersey, 610

Taylor Road, Piscataway, NJ 08854, USA.

The w3DNA (web 3DNA) server is a user-friendly web-based interface to the 3DNA

suite of programs for the analysis, reconstruction, and visualization of

three-dimensional (3D) nucleic-acid-containing structures, including their

complexes with proteins and other ligands. The server allows the user to

determine a wide variety of conformational parameters in a given structure--such

as the identities and rigid-body parameters of interacting nucleic-acid bases and

base-pair steps, the nucleotides comprising helical fragments, etc. It is also

possible to build 3D models of arbitrary nucleotide sequences and helical types,

customized single-stranded and double-helical structures with user-defined

base-pair parameters and sequences, and models of DNA 'decorated' at user-defined

sites with proteins and other molecules. The visualization component offers

unique, publication-quality representations of nucleic-acid structures, such as

'block' images of bases and base pairs and stacking diagrams of interacting

nucleotides. The w3DNA web server, located at http://w3dna.rutgers.edu, is free

and open to all users with no login requirement.

DOI: 10.1093/nar/gkp358

PMCID: PMC2703980

PMID: 19474339 [Indexed for MEDLINE]

2186. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W452-8. doi: 10.1093/nar/gkp409.

Epub 2009 May 27.

PLecDom: a program for identification and analysis of plant lectin domains.

Shridhar S(1), Chattopadhyay D, Yadav G.

Author information:

(1)Computational Biology Laboratory, National Institute of Plant Genome Research,

Aruna Asaf Ali Marg, New Delhi 110067, India.

PLecDom is a program for detection of Plant Lectin Domains in a polypeptide or

EST sequence, followed by a classification of the identified domains into known

families. The web server is a collection of plant lectin domain families

represented by alignments and profile Hidden Markov Models. PLecDom was developed

after a rigorous analysis of evolutionary relationships between available

sequences of lectin domains with known specificities. Users can test their

sequences for potential lectin domains, catalog the identified domains into broad

substrate classes, estimate the extent of divergence of new domains with existing

homologs, extract domain boundaries and examine flanking sequences for further

analysis. The high prediction accuracy of PLecDom combined with the ease with

which it handles large scale input, enabled us to apply the program to protein

and EST data from 48 plant genome-sequencing projects in various stages of

completion. Our results represent a significant enrichment of the currently

annotated plant lectins, and highlight potential targets for biochemical

characterization. The search algorithm requires input in fasta format and is

designed to process simultaneous connection requests from multiple users, such

that huge sets of input sequences can be scanned in a matter of seconds. PLecDom

is available at http://www.nipgr.res.in/plecdom.html.

DOI: 10.1093/nar/gkp409

PMCID: PMC2703983

PMID: 19474338 [Indexed for MEDLINE]

2187. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W266-72. doi:

10.1093/nar/gkp412. Epub 2009 May 25.

MirZ: an integrated microRNA expression atlas and target prediction resource.

Hausser J(1), Berninger P, Rodak C, Jantscher Y, Wirth S, Zavolan M.

Author information:

(1)Biozentrum, Universität Basel and Swiss Institute of Bioinformatics,

Klingelbergstrasse 50-70, 4056 Basel, Switzerland.

MicroRNAs (miRNAs) are short RNAs that act as guides for the degradation and

translational repression of protein-coding mRNAs. A large body of work showed

that miRNAs are involved in the regulation of a broad range of biological

functions, from development to cardiac and immune system function, to metabolism,

to cancer. For most of the over 500 miRNAs that are encoded in the human genome

the functions still remain to be uncovered. Identifying miRNAs whose expression

changes between cell types or between normal and pathological conditions is an

important step towards characterizing their function as is the prediction of

mRNAs that could be targeted by these miRNAs. To provide the community the

possibility of exploring interactively miRNA expression patterns and the

candidate targets of miRNAs in an integrated environment, we developed the MirZ

web server, which is accessible at www.mirz.unibas.ch. The server provides

experimental and computational biologists with statistical analysis and data

mining tools operating on up-to-date databases of sequencing-based miRNA

expression profiles and of predicted miRNA target sites in species ranging from

Caenorhabditis elegans to Homo sapiens.

DOI: 10.1093/nar/gkp412

PMCID: PMC2703880

PMID: 19468042 [Indexed for MEDLINE]

2188. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W376-83. doi:

10.1093/nar/gkp410. Epub 2009 May 21.

HotSpot Wizard: a web server for identification of hot spots in protein

engineering.

Pavelka A(1), Chovancova E, Damborsky J.

Author information:

(1)Loschmidt Laboratories, Institute of Experimental Biology and National Centre

for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5/A4,

625 00 Brno, Czech Republic.

HotSpot Wizard is a web server for automatic identification of 'hot spots' for

engineering of substrate specificity, activity or enantioselectivity of enzymes

and for annotation of protein structures. The web server implements the protein

engineering protocol, which targets evolutionarily variable amino acid positions

located in the active site or lining the access tunnels. The 'hot spots' for

mutagenesis are selected through the integration of structural, functional and

evolutionary information obtained from: (i) the databases RCSB PDB, UniProt,

PDBSWS, Catalytic Site Atlas and nr NCBI and (ii) the tools CASTp, CAVER, BLAST,

CD-HIT, MUSCLE and Rate4Site. The protein structure and e-mail address are the

only obligatory inputs for the calculation. In the output, HotSpot Wizard lists

annotated residues ordered by estimated mutability. The results of the analysis

are mapped on the enzyme structure and visualized in the web browser using Jmol.

The HotSpot Wizard server should be useful for protein engineers interested in

exploring the structure of their favourite protein and for the design of

mutations in site-directed mutagenesis and focused directed evolution

experiments. HotSpot Wizard is available at

http://loschmidt.chemi.muni.cz/hotspotwizard/.

DOI: 10.1093/nar/gkp410

PMCID: PMC2703904

PMID: 19465397 [Indexed for MEDLINE]

2189. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W95-W100. doi:

10.1093/nar/gkp377. Epub 2009 May 21.

primers4clades: a web server that uses phylogenetic trees to design

lineage-specific PCR primers for metagenomic and diversity studies.

Contreras-Moreira B(1), Sachman-Ruiz B, Figueroa-Palacios I, Vinuesa P.

Author information:

(1)Estación Experimental de Aula Dei, Consejo Superior de Investigaciones

Científicas, Zaragoza, Mexico.

Primers4clades is an easy-to-use web server that implements a fully automatic PCR

primer design pipeline for cross-species amplification of novel sequences from

metagenomic DNA, or from uncharacterized organisms, belonging to user-specified

phylogenetic clades or taxa. The server takes a set of non-aligned protein coding

genes, with or without introns, aligns them and computes a neighbor-joining tree,

which is displayed on screen for easy selection of species or sequence clusters

to design lineage-specific PCR primers. Primers4clades implements an extended

CODEHOP primer design strategy based on both DNA and protein multiple sequence

alignments. It evaluates several thermodynamic properties of the oligonucleotide

pairs, and computes the phylogenetic information content of the predicted

amplicon sets from Shimodaira-Hasegawa-like branch support values of maximum

likelihood phylogenies. A non-redundant set of primer formulations is returned,

ranked according to their thermodynamic properties. An amplicon distribution map

provides a convenient overview of the coverage of the target locus. Altogether

these features greatly help the user in making an informed choice between

alternative primer pair formulations. Primers4clades is available at two mirror

sites: http://maya.ccg.unam.mx/primers4clades/and

http://floresta.eead.csic.es/primers4clades/. Three demo data sets and a

comprehensive documentation/tutorial page are provided for easy testing of the

server's capabilities and interface.

DOI: 10.1093/nar/gkp377

PMCID: PMC2703966

PMID: 19465390 [Indexed for MEDLINE]

2190. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W417-21. doi:

10.1093/nar/gkp329. Epub 2009 May 21.

d-Omix: a mixer of generic protein domain analysis tools.

Wichadakul D(1), Numnark S, Ingsriswang S.

Author information:

(1)National Center for Genetic Engineering and Biotechnology (BIOTEC) -

Information Systems Laboratory, Pathumthani, Thailand. duangdao.wic@biotec.or.th

Domain combination provides important clues to the roles of protein domains in

protein function, interaction and evolution. We have developed a web server

d-Omix (a Mixer of Protein Domain Analysis Tools) aiming as a unified platform to

analyze, compare and visualize protein data sets in various aspects of protein

domain combinations. With InterProScan files for protein sets of interest

provided by users, the server incorporates four services for domain analyses.

First, it constructs protein phylogenetic tree based on a distance matrix

calculated from protein domain architectures (DAs), allowing the comparison with

a sequence-based tree. Second, it calculates and visualizes the versatility,

abundance and co-presence of protein domains via a domain graph. Third, it

compares the similarity of proteins based on DA alignment. Fourth, it builds a

putative protein network derived from domain-domain interactions from DOMINE.

Users may select a variety of input data files and flexibly choose domain search

tools (e.g. hmmpfam, superfamily) for a specific analysis. Results from the

d-Omix could be interactively explored and exported into various formats such as

SVG, JPG, BMP and CSV. Users with only protein sequences could prepare an

InterProScan file using a service provided by the server as well. The d-Omix web

server is freely available at http://www.biotec.or.th/isl/Domix.

DOI: 10.1093/nar/gkp329

PMCID: PMC2703976

PMID: 19465389 [Indexed for MEDLINE]

2191. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W317-22. doi:

10.1093/nar/gkp416. Epub 2009 May 22.

GeneCodis: interpreting gene lists through enrichment analysis and integration of

diverse biological information.

Nogales-Cadenas R(1), Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F,

Carazo JM, Pascual-Montano A.

Author information:

(1)Computer Architecture Department, Complutense University of Madrid, Madrid,

Spain.

GeneCodis is a web server application for functional analysis of gene lists that

integrates different sources of information and finds modular patterns of

interrelated annotations. This integrative approach has proved to be useful for

the interpretation of high-throughput experiments and therefore a new version of

the system has been developed to expand its functionality and scope. GeneCodis

now expands the functional information with regulatory patterns and user-defined

annotations, offering the possibility of integrating all sources of information

in the same analysis. Traditional singular enrichment is now permitted and more

organisms and gene identifiers have been added to the database. The application

has been re-engineered to improve performance, accessibility and scalability. In

addition, GeneCodis can now be accessed through a public SOAP web services

interface, enabling users to perform analysis from their own scripts and

workflows. The application is freely available at http://genecodis.dacya.ucm.es.

DOI: 10.1093/nar/gkp416

PMCID: PMC2703901

PMID: 19465387 [Indexed for MEDLINE]

2192. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W539-44. doi:

10.1093/nar/gkp411. Epub 2009 May 22.

COPS--a novel workbench for explorations in fold space.

Suhrer SJ(1), Wiederstein M, Gruber M, Sippl MJ.

Author information:

(1)Center of Applied Molecular Engineering, Division of Bioinformatics,

University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria.

The COPS (Classification Of Protein Structures) web server provides access to the

complete repertoire of known protein structures and protein structural domains.

The COPS classification encodes pairwise structural similarities as quantified

metric relationships. The resulting metrical structure is mapped to a

hierarchical tree, which is largely equivalent to the structure of a file

browser. Exploiting this relationship we implemented the Fold Space Navigator, a

tool that makes navigation in fold space as convenient as browsing through a file

system. Moreover, pairwise structural similarities among the domains can be

visualized and inspected instantaneously. COPS is updated weekly and stays

concurrent with the PDB repository. The server also exposes the COPS

classification pipeline. Newly determined structures uploaded to the server are

chopped into domains, the locations of the new domains in the classification tree

are determined, and their neighborhood can be immediately explored through the

Fold Space Navigator. The COPS web server is accessible at

http://cops.services.came.sbg.ac.at/.

DOI: 10.1093/nar/gkp411

PMCID: PMC2703906

PMID: 19465386 [Indexed for MEDLINE]

2193. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W565-70. doi:

10.1093/nar/gkp405. Epub 2009 May 22.

MolLoc: a web tool for the local structural alignment of molecular surfaces.

Angaran S(1), Bock ME, Garutti C, Guerra C.

Author information:

(1)Department of Information Engineering, University of Padova, Via Gradenigo 6a,

Padova, Italy.

MolLoc stands for Molecular Local surface comparison, and is a web server for the

structural comparison of molecular surfaces. Given two structures in PDB format,

the user can compare their binding sites, cavities or any arbitrary residue

selection. Moreover, the web server allows the comparison of a query structure

with a list of structures. Each comparison produces a structural alignment that

maximizes the extension of the superimposition of the surfaces, and returns the

pairs of atoms with similar physicochemical properties that are close in space

after the superimposition. Based on this subset of atoms sharing similar

physicochemical properties a new rototranslation is derived that best

superimposes them. MolLoc approach is both local and surface-oriented, and

therefore it can be particularly useful when testing if molecules with different

sequences and folds share any local surface similarity. The MolLoc web server is

available at http://bcb.dei.unipd.it/MolLoc.

DOI: 10.1093/nar/gkp405

PMCID: PMC2703929

PMID: 19465382 [Indexed for MEDLINE]

2194. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W480-4. doi: 10.1093/nar/gkp431.

Epub 2009 May 22.

The SALAMI protein structure search server.

Margraf T(1), Schenk G, Torda AE.

Author information:

(1)Centre for Bioinformatics, University of Hamburg, Bundesstr. 43, 20146

Hamburg, Germany. margraf@zbh.uni-hamburg.de

Protein structures often show similarities to another which would not be seen at

the sequence level. Given the coordinates of a protein chain, the SALAMI server

at www.zbh.uni-hamburg.de/salami will search the protein data bank and return a

set of similar structures without using sequence information. The results page

lists the related proteins, details of the sequence and structure similarity and

implied sequence alignments. Via a simple structure viewer, one can view

superpositions of query and library structures and finally download superimposed

coordinates. The alignment method is very tolerant of large gaps and insertions,

and tends to produce slightly longer alignments than other similar programs.

DOI: 10.1093/nar/gkp431

PMCID: PMC2703935

PMID: 19465380 [Indexed for MEDLINE]

2195. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W575-80. doi:

10.1093/nar/gkp418. Epub 2009 May 22.

RHYTHM--a server to predict the orientation of transmembrane helices in channels

and membrane-coils.

Rose A(1), Lorenzen S, Goede A, Gruening B, Hildebrand PW.

Author information:

(1)Institute for Medical Physics and Biophysics, Charité, University Medicine

Berlin, Ziegelstrasse 5-9, 10098 Berlin, Germany.

RHYTHM is a web server that predicts buried versus exposed residues of helical

membrane proteins. Starting from a given protein sequence, secondary and tertiary

structure information is calculated by RHYTHM within only a few seconds. The

prediction applies structural information from a growing data base of

precalculated packing files and evolutionary information from sequence patterns

conserved in a representative dataset of membrane proteins ('Pfam-domains'). The

program uses two types of position specific matrices to account for the different

geometries of packing in channels and transporters ('channels') or other membrane

proteins ('membrane-coils'). The output provides information on the secondary

structure and topology of the protein and specifically on the contact type of

each residue and its conservation. This information can be downloaded as a

graphical file for illustration, a text file for analysis and statistics and a

PyMOL file for modeling purposes. The server can be freely accessed at: URL:

http://proteinformatics.de/rhythm.

DOI: 10.1093/nar/gkp418

PMCID: PMC2703963

PMID: 19465378 [Indexed for MEDLINE]

2196. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W612-6. doi: 10.1093/nar/gkp417.

Epub 2009 May 22.

SEPPA: a computational server for spatial epitope prediction of protein antigens.

Sun J(1), Wu D, Xu T, Wang X, Xu X, Tao L, Li YX, Cao ZW.

Author information:

(1)Department of Biomedical Engineering, School of Life Sciences and Technology,

Tongji University, Shanghai 200092, China.

In recent years, a lot of efforts have been made in conformational epitope

prediction as antigen proteins usually bind antibodies with an assembly of

sequentially discontinuous and structurally compact surface residues. Currently,

only a few methods for spatial epitope prediction are available with focus on

single residue propensity scales or continual segments clustering. In the method

of SEPPA, a concept of 'unit patch of residue triangle' was introduced to better

describe the local spatial context in protein surface. Besides that, SEPPA

incorporated clustering coefficient to describe the spatial compactness of

surface residues. Validated by independent testing datasets, SEPPA gave an

average AUC value over 0.742 and produced a successful pick-up rate of 96.64%.

Comparing with peers, SEPPA shows significant improvement over other popular

methods like CEP, DiscoTope and BEpro. In addition, the threshold scores for

certain accuracy, sensitivity and specificity are provided online to give the

confidence level of the spatial epitope identification. The web server can be

accessed at http://lifecenter.sgst.cn/seppa/index.php. Batch query is supported.

DOI: 10.1093/nar/gkp417

PMCID: PMC2703964

PMID: 19465377 [Indexed for MEDLINE]

2197. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W153-9. doi: 10.1093/nar/gkp392.

Epub 2009 May 20.

SENT: semantic features in text.

Vazquez M(1), Carmona-Saez P, Nogales-Cadenas R, Chagoyen M, Tirado F, Carazo JM,

Pascual-Montano A.

Author information:

(1)Software Engineering Department, Complutense University and Biocomputing Unit,

National Center for Biotechnology, CNB-CSIC, Madrid, Spain.

We present SENT (semantic features in text), a functional interpretation tool

based on literature analysis. SENT uses Non-negative Matrix Factorization to

identify topics in the scientific articles related to a collection of genes or

their products, and use them to group and summarize these genes. In addition, the

application allows users to rank and explore the articles that best relate to the

topics found, helping put the analysis results into context. This approach is

useful as an exploratory step in the workflow of interpreting and understanding

experimental data, shedding some light into the complex underlying biological

mechanisms. This tool provides a user-friendly interface via a web site, and a

programmatic access via a SOAP web server. SENT is freely accessible at

http://sent.dacya.ucm.es.

DOI: 10.1093/nar/gkp392

PMCID: PMC2703940

PMID: 19458159 [Indexed for MEDLINE]

2198. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W202-8. doi: 10.1093/nar/gkp335.

Epub 2009 May 20.

MEME SUITE: tools for motif discovery and searching.

Bailey TL(1), Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW,

Noble WS.

Author information:

(1)Institute for Molecular Bioscience, University of Queensland, Brisbane,

Queensland, Australia. t.bailey@imb.uq.edu.au

The MEME Suite web server provides a unified portal for online discovery and

analysis of sequence motifs representing features such as DNA binding sites and

protein interaction domains. The popular MEME motif discovery algorithm is now

complemented by the GLAM2 algorithm which allows discovery of motifs containing

gaps. Three sequence scanning algorithms--MAST, FIMO and GLAM2SCAN--allow

scanning numerous DNA and protein sequence databases for motifs discovered by

MEME and GLAM2. Transcription factor motifs (including those discovered using

MEME) can be compared with motifs in many popular motif databases using the motif

database scanning algorithm TOMTOM. Transcription factor motifs can be further

analyzed for putative function by association with Gene Ontology (GO) terms using

the motif-GO term association tool GOMO. MEME output now contains sequence LOGOS

for each discovered motif, as well as buttons to allow motifs to be conveniently

submitted to the sequence and motif database scanning algorithms (MAST, FIMO and

TOMTOM), or to GOMO, for further analysis. GLAM2 output similarly contains

buttons for further analysis using GLAM2SCAN and for rerunning GLAM2 with

different parameters. All of the motif-based tools are now implemented as web

services via Opal. Source code, binaries and a web server are freely available

for noncommercial use at http://meme.nbcr.net.

DOI: 10.1093/nar/gkp335

PMCID: PMC2703892

PMID: 19458158 [Indexed for MEDLINE]

2199. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W474-9. doi: 10.1093/nar/gkp387.

Epub 2009 May 20.

RosettaAntibody: antibody variable region homology modeling server.

Sircar A(1), Kim ET, Gray JJ.

Author information:

(1)Department of Chemical and Biomolecular Engineering, Johns Hopkins University,

3400 N. Charles Street, Baltimore, MD 21218, USA.

The RosettaAntibody server (http://antibody.graylab.jhu.edu) predicts the

structure of an antibody variable region given the amino-acid sequences of the

respective light and heavy chains. In an initial stage, the server identifies and

displays the most sequence homologous template structures for the light and heavy

framework regions and each of the complementarity determining region (CDR) loops.

Subsequently, the most homologous templates are assembled into a side-chain

optimized crude model, and the server returns a picture and coordinate file. For

users requesting a high-resolution model, the server executes the full

RosettaAntibody protocol which additionally models the hyper-variable CDR H3

loop. The high-resolution protocol also relieves steric clashes by optimizing the

CDR backbone torsion angles and by simultaneously perturbing the relative

orientation of the light and heavy chains. RosettaAntibody generates 2000

independent structures, and the server returns pictures, coordinate files, and

detailed scoring information for the 10 top-scoring models. The 10 models enable

users to use rational judgment in choosing the best model or to use the set as an

ensemble for further studies such as docking. The high-resolution models

generated by RosettaAntibody have been used for the successful prediction of

antibody-antigen complex structures.

DOI: 10.1093/nar/gkp387

PMCID: PMC2703951

PMID: 19458157 [Indexed for MEDLINE]

2200. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W179-84. doi:

10.1093/nar/gkp370. Epub 2009 May 13.

PhyloPars: estimation of missing parameter values using phylogeny.

Bruggeman J(1), Heringa J, Brandt BW.

Author information:

(1)Department of Theoretical Biology, Faculty of Earth & Life Sciences, VU

University Amsterdam, The Netherlands. jorn.bruggeman@falw.vu.nl

A wealth of information on metabolic parameters of a species can be inferred from

observations on species that are phylogenetically related. Phylogeny-based

information can complement direct empirical evidence, and is particularly

valuable if experiments on the species of interest are not feasible. The

PhyloPars web server provides a statistically consistent method that combines an

incomplete set of empirical observations with the species phylogeny to produce a

complete set of parameter estimates for all species. It builds upon a

state-of-the-art evolutionary model, extended with the ability to handle missing

data. The resulting approach makes optimal use of all available information to

produce estimates that can be an order of magnitude more accurate than ad-hoc

alternatives. Uploading a phylogeny and incomplete feature matrix suffices to

obtain estimates of all missing values, along with a measure of certainty.

Real-time cross-validation provides further insight in the accuracy and bias

expected for estimated values. The server allows for easy, efficient estimation

of metabolic parameters, which can benefit a wide range of fields including

systems biology and ecology. PhyloPars is available at:

http://www.ibi.vu.nl/programs/phylopars/.

DOI: 10.1093/nar/gkp370

PMCID: PMC2703881

PMID: 19443453 [Indexed for MEDLINE]

2201. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W390-5. doi: 10.1093/nar/gkp339.

Epub 2009 May 13.

INTREPID: a web server for prediction of functionally important residues by

evolutionary analysis.

Sankararaman S(1), Kolaczkowski B, Sjölander K.

Author information:

(1)Department of Computer Science, University of California, Berkeley, USA.

We present the INTREPID web server for predicting functionally important residues

in proteins. INTREPID has been shown to boost the recall and precision of

catalytic residue prediction over other sequence-based methods and can be used to

identify other types of functional residues. The web server takes an input

protein sequence, gathers homologs, constructs a multiple sequence alignment and

phylogenetic tree and finally runs the INTREPID method to assign a score to each

position. Residues predicted to be functionally important are displayed on

homologous 3D structures (where available), highlighting spatial patterns of

conservation at various significance thresholds. The INTREPID web server is

available at http://phylogenomics.berkeley.edu/intrepid.

DOI: 10.1093/nar/gkp339

PMCID: PMC2703888

PMID: 19443452 [Indexed for MEDLINE]

2202. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W485-91. doi:

10.1093/nar/gkp368. Epub 2009 May 13.

@TOME-2: a new pipeline for comparative modeling of protein-ligand complexes.

Pons JL(1), Labesse G.

Author information:

(1)A.B.C.I.S., CNRS UMR5048-Universités Montpellier 1 & Montpellier II and INSERM

U554, 29, Rue de Navacelles, 34090 Montpellier Cedex, France.

@TOME 2.0 is new web pipeline dedicated to protein structure modeling and small

ligand docking based on comparative analyses. @TOME 2.0 allows fold recognition,

template selection, structural alignment editing, structure comparisons, 3D-model

building and evaluation. These tasks are routinely used in sequence analyses for

structure prediction. In our pipeline the necessary software is efficiently

interconnected in an original manner to accelerate all the processes.

Furthermore, we have also connected comparative docking of small ligands that is

performed using protein-protein superposition. The input is a simple protein

sequence in one-letter code with no comment. The resulting 3D model,

protein-ligand complexes and structural alignments can be visualized through

dedicated Web interfaces or can be downloaded for further studies. These original

features will aid in the functional annotation of proteins and the selection of

templates for molecular modeling and virtual screening. Several examples are

described to highlight some of the new functionalities provided by this pipeline.

The server and its documentation are freely available at

http://abcis.cbs.cnrs.fr/AT2/

DOI: 10.1093/nar/gkp368

PMCID: PMC2703933

PMID: 19443448 [Indexed for MEDLINE]

2203. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W84-9. doi: 10.1093/nar/gkp373.

Epub 2009 May 12.

Berkeley PHOG: PhyloFacts orthology group prediction web server.

Datta RS(1), Meacham C, Samad B, Neyer C, Sjölander K.

Author information:

(1)QB3 Institute, University of California, Berkeley, CA, USA.

ruchira@berkeley.edu

Ortholog detection is essential in functional annotation of genomes, with

applications to phylogenetic tree construction, prediction of protein-protein

interaction and other bioinformatics tasks. We present here the PHOG web server

employing a novel algorithm to identify orthologs based on phylogenetic analysis.

Results on a benchmark dataset from the TreeFam-A manually curated orthology

database show that PHOG provides a combination of high recall and precision

competitive with both InParanoid and OrthoMCL, and allows users to target

different taxonomic distances and precision levels through the use of

tree-distance thresholds. For instance, OrthoMCL-DB achieved 76% recall and 66%

precision on this dataset; at a slightly higher precision (68%) PHOG achieves 10%

higher recall (86%). InParanoid achieved 87% recall at 24% precision on this

dataset, while a PHOG variant designed for high recall achieves 88% recall at 61%

precision, increasing precision by 37% over InParanoid. PHOG is based on

pre-computed trees in the PhyloFacts resource, and contains over 366 K orthology

groups with a minimum of three species. Predicted orthologs are linked to GO

annotations, pathway information and biological literature. The PHOG web server

is available at http://phylofacts.berkeley.edu/orthologs/.

DOI: 10.1093/nar/gkp373

PMCID: PMC2703887

PMID: 19435885 [Indexed for MEDLINE]

2204. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W90-4. doi: 10.1093/nar/gkp360.

Epub 2009 May 12.

COMPASS server for homology detection: improved statistical accuracy, speed and

functionality.

Sadreyev RI(1), Tang M, Kim BH, Grishin NV.

Author information:

(1)Howard Hughes Medical Institute, University of Texas Southwestern Medical

Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA.

sadreyev@chop.swmed.edu

COMPASS is a profile-based method for the detection of remote sequence similarity

and the prediction of protein structure. Here we describe a recently improved

public web server of COMPASS, http://prodata.swmed.edu/compass. The server

features three major developments: (i) improved statistical accuracy; (ii)

increased speed from parallel implementation; and (iii) new functional features

facilitating structure prediction. These features include visualization tools

that allow the user to quickly and effectively analyze specific local structural

region predictions suggested by COMPASS alignments. As an application example, we

describe the structural, evolutionary and functional analysis of a protein with

unknown function that served as a target in the recent CASP8 (Critical Assessment

of Techniques for Protein Structure Prediction round 8). URL:

http://prodata.swmed.edu/compass.

DOI: 10.1093/nar/gkp360

PMCID: PMC2703893

PMID: 19435884 [Indexed for MEDLINE]

2205. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W277-80. doi:

10.1093/nar/gkp367. Epub 2009 May 12.

CENTROIDFOLD: a web server for RNA secondary structure prediction.

Sato K(1), Hamada M, Asai K, Mituyama T.

Author information:

(1)Japan Biological Informatics Consortium, 2-45 Aomi, Koto-ku, Tokyo 135-8073,

Japan. sato-kengo@aist.go.jp

The CENTROIDFOLD web server (http://www.ncrna.org/centroidfold/) is a web

application for RNA secondary structure prediction powered by one of the most

accurate prediction engine. The server accepts two kinds of sequence data: a

single RNA sequence and a multiple alignment of RNA sequences. It responses with

a prediction result shown as a popular base-pair notation and a graph

representation. PDF version of the graph representation is also available. For a

multiple alignment sequence, the server predicts a common secondary structure.

Usage of the server is quite simple. You can paste a single RNA sequence (FASTA

or plain sequence text) or a multiple alignment (CLUSTAL-W format) into the

textarea then click on the 'execute CentroidFold' button. The server quickly

responses with a prediction result. The major advantage of this server is that it

employs our original CentroidFold software as its prediction engine which scores

the best accuracy in our benchmark results. Our web server is freely available

with no login requirement.

DOI: 10.1093/nar/gkp367

PMCID: PMC2703931

PMID: 19435882 [Indexed for MEDLINE]

2206. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W68-76. doi: 10.1093/nar/gkp347.

Epub 2009 May 11.

miRanalyzer: a microRNA detection and analysis tool for next-generation

sequencing experiments.

Hackenberg M(1), Sturm M, Langenberger D, Falcón-Pérez JM, Aransay AM.

Author information:

(1)Functional Genomics Unit, CIC bioGUNE, CIBERehd, Technology Park of Bizkaia,

48160 Derio, Bizkaia, Spain.

Next-generation sequencing allows now the sequencing of small RNA molecules and

the estimation of their expression levels. Consequently, there will be a high

demand of bioinformatics tools to cope with the several gigabytes of sequence

data generated in each single deep-sequencing experiment. Given this scene, we

developed miRanalyzer, a web server tool for the analysis of deep-sequencing

experiments for small RNAs. The web server tool requires a simple input file

containing a list of unique reads and its copy numbers (expression levels). Using

these data, miRanalyzer (i) detects all known microRNA sequences annotated in

miRBase, (ii) finds all perfect matches against other libraries of transcribed

sequences and (iii) predicts new microRNAs. The prediction of new microRNAs is an

especially important point as there are many species with very few known

microRNAs. Therefore, we implemented a highly accurate machine learning algorithm

for the prediction of new microRNAs that reaches AUC values of 97.9% and recall

values of up to 75% on unseen data. The web tool summarizes all the described

steps in a single output page, which provides a comprehensive overview of the

analysis, adding links to more detailed output pages for each analysis module.

miRanalyzer is available at http://web.bioinformatics.cicbiogune.es/microRNA/.

DOI: 10.1093/nar/gkp347

PMCID: PMC2703919

PMID: 19433510 [Indexed for MEDLINE]

2207. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W559-64. doi:

10.1093/nar/gkp359. Epub 2009 May 11.

SLITHER: a web server for generating contiguous conformations of substrate

molecules entering into deep active sites of proteins or migrating through

channels in membrane transporters.

Lee PH(1), Kuo KL, Chu PY, Liu EM, Lin JH.

Author information:

(1)Division of Mechanics, Research Center for Applied Sciences, Academia Sinica,

Taipei, Taiwan.

Many proteins use a long channel to guide the substrate or ligand molecules into

the well-defined active sites for catalytic reactions or for switching molecular

states. In addition, substrates of membrane transporters can migrate to another

side of cellular compartment by means of certain selective mechanisms. SLITHER

(http://bioinfo.mc.ntu.edu.tw/slither/or http://slither.rcas.sinica.edu.tw/) is a

web server that can generate contiguous conformations of a molecule along a

curved tunnel inside a protein, and the binding free energy profile along the

predicted channel pathway. SLITHER adopts an iterative docking scheme, which

combines with a puddle-skimming procedure, i.e. repeatedly elevating the

potential energies of the identified global minima, thereby determines the

contiguous binding modes of substrates inside the protein. In contrast to some

programs that are widely used to determine the geometric dimensions in the ion

channels, SLITHER can be applied to predict whether a substrate molecule can

crawl through an inner channel or a half-channel of proteins across surmountable

energy barriers. Besides, SLITHER also provides the list of the pore-facing

residues, which can be directly compared with many genetic diseases. Finally, the

adjacent binding poses determined by SLITHER can also be used for fragment-based

drug design.

DOI: 10.1093/nar/gkp359

PMCID: PMC2703944

PMID: 19433508 [Indexed for MEDLINE]

2208. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W652-60. doi:

10.1093/nar/gkp356. Epub 2009 May 8.

MetaboAnalyst: a web server for metabolomic data analysis and interpretation.

Xia J(1), Psychogios N, Young N, Wishart DS.

Author information:

(1)Department of Biological Sciences, University of Alberta, Edmonton AB T6G 2E8,

Canada.

Metabolomics is a newly emerging field of 'omics' research that is concerned with

characterizing large numbers of metabolites using NMR, chromatography and mass

spectrometry. It is frequently used in biomarker identification and the metabolic

profiling of cells, tissues or organisms. The data processing challenges in

metabolomics are quite unique and often require specialized (or expensive) data

analysis software and a detailed knowledge of cheminformatics, bioinformatics and

statistics. In an effort to simplify metabolomic data analysis while at the same

time improving user accessibility, we have developed a freely accessible,

easy-to-use web server for metabolomic data analysis called MetaboAnalyst.

Fundamentally, MetaboAnalyst is a web-based metabolomic data processing tool not

unlike many of today's web-based microarray analysis packages. It accepts a

variety of input data (NMR peak lists, binned spectra, MS peak lists,

compound/concentration data) in a wide variety of formats. It also offers a

number of options for metabolomic data processing, data normalization,

multivariate statistical analysis, graphing, metabolite identification and

pathway mapping. In particular, MetaboAnalyst supports such techniques as: fold

change analysis, t-tests, PCA, PLS-DA, hierarchical clustering and a number of

more sophisticated statistical or machine learning methods. It also employs a

large library of reference spectra to facilitate compound identification from

most kinds of input spectra. MetaboAnalyst guides users through a step-by-step

analysis pipeline using a variety of menus, information hyperlinks and check

boxes. Upon completion, the server generates a detailed report describing each

method used, embedded with graphical and tabular outputs. MetaboAnalyst is

capable of handling most kinds of metabolomic data and was designed to perform

most of the common kinds of metabolomic data analyses. MetaboAnalyst is

accessible at http://www.metaboanalyst.ca.

DOI: 10.1093/nar/gkp356

PMCID: PMC2703878

PMID: 19429898 [Indexed for MEDLINE]

2209. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W571-4. doi: 10.1093/nar/gkp338.

Epub 2009 May 8.

SuperLooper--a prediction server for the modeling of loops in globular and

membrane proteins.

Hildebrand PW(1), Goede A, Bauer RA, Gruening B, Ismer J, Michalsky E, Preissner

R.

Author information:

(1)Institute of Medical Physics and Biophysics, Charité, University of Medicine,

Berlin, Germany. peter.hildebrand@charite.de

SuperLooper provides the first online interface for the automatic, quick and

interactive search and placement of loops in proteins (LIP). A database

containing half a billion segments of water-soluble proteins with lengths up to

35 residues can be screened for candidate loops. A specified database containing

180,000 membrane loops in proteins (LIMP) can be searched, alternatively. Loop

candidates are scored based on sequence criteria and the root mean square

deviation (RMSD) of the stem atoms. Searching LIP, the average global RMSD of the

respective top-ranked loops to the original loops is benchmarked to be <2 A, for

loops up to six residues or <3 A for loops shorter than 10 residues. Other

suitable conformations may be selected and directly visualized on the web server

from a top-50 list. For user guidance, the sequence homology between the template

and the original sequence, proline or glycine exchanges or close contacts between

a loop candidate and the remainder of the protein are denoted. For membrane

proteins, the expansions of the lipid bilayer are automatically modeled using the

TMDET algorithm. This allows the user to select the optimal membrane protein loop

concerning its relative orientation to the lipid bilayer. The server is online

since October 2007 and can be freely accessed at URL:

http://bioinformatics.charite.de/superlooper/.

DOI: 10.1093/nar/gkp338

PMCID: PMC2703960

PMID: 19429894 [Indexed for MEDLINE]

2210. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W465-8. doi: 10.1093/nar/gkp363.

Epub 2009 May 8.

TOPCONS: consensus prediction of membrane protein topology.

Bernsel A(1), Viklund H, Hennerdal A, Elofsson A.

Author information:

(1)Center for Biomembrane Research, Department of Biochemistry and Biophysics,

Stockholm University, SE-106 91, Sweden.

TOPCONS (http://topcons.net/) is a web server for consensus prediction of

membrane protein topology. The underlying algorithm combines an arbitrary number

of topology predictions into one consensus prediction and quantifies the

reliability of the prediction based on the level of agreement between the

underlying methods, both on the protein level and on the level of individual TM

regions. Benchmarking the method shows that overall performance levels match the

best available topology prediction methods, and for sequences with high

reliability scores, performance is increased by approximately 10 percentage

points. The web interface allows for constraining parts of the sequence to a

known inside/outside location, and detailed results are displayed both

graphically and in text format.

DOI: 10.1093/nar/gkp363

PMCID: PMC2703981

PMID: 19429891 [Indexed for MEDLINE]

2211. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W363-8. doi: 10.1093/nar/gkp299.

Epub 2009 May 8.

ClanTox: a classifier of short animal toxins.

Naamati G(1), Askenazi M, Linial M.

Author information:

(1)Computer Science and Engineering, The Hebrew University of Jerusalem, Israel.

Toxins are detected in sporadic species along the evolutionary tree of the animal

kingdom. Venomous animals include scorpions, snakes, bees, wasps, frogs and

numerous animals living in the sea such as the stonefish, snail, jellyfish, hydra

and more. Interestingly, proteins that share a common scaffold with animal toxins

also exist in non-venomous species. However, due to their short length and

primary sequence diversity, these, toxin-like proteins remain undetected by

classical search engines and genome annotation tools. We construct a toxin

classification machine and web server called ClanTox (Classifier of Animal

Toxins) that is based on the extraction of sequence-driven features from the

primary protein sequence followed by the application of a classification system

trained on known animal toxins. For a given input list of sequences, from

venomous or non-venomous settings, the ClanTox system predicts whether each

sequence is toxin-like. ClanTox provides a ranked list of positively predicted

candidates according to statistical confidence. For each protein, additional

information is presented including the presence of a signal peptide, the number

of cysteine residues and the associated functional annotations. ClanTox is a

discovery-prediction tool for a relatively overlooked niche of toxin-like cell

modulators, many of which are therapeutic agent candidates. The ClanTox web

server is freely accessible at http://www.clantox.cs.huji.ac.il.

DOI: 10.1093/nar/gkp299

PMCID: PMC2703885

PMID: 19429697 [Indexed for MEDLINE]

2212. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W53-6. doi: 10.1093/nar/gkp301.

Epub 2009 May 8.

SIB-BLAST: a web server for improved delineation of true and false positives in

PSI-BLAST searches.

Lee MM(1), Chan MK, Bundschuh R.

Author information:

(1)The Ohio State Biophysics Program, Ohio State University, Columbus, OH

43210-1117, USA.

A SIB-BLAST web server (http://sib-blast.osc.edu) has been established for

investigators to use the SimpleIsBeautiful (SIB) algorithm for sequence-based

homology detection. SIB was developed to overcome the model corruption frequently

observed in the later iterations of PSI-BLAST searches. The algorithm compares

resultant hits from the second iteration to the final iteration of a PSI-BLAST

search, calculates the figure of merit for each 'overlapped' hit and re-ranks the

hits according to their figure of merit. By validating hits generated from the

last profile against hits from the first profile when the model is least

corrupted, the true and false positives are better delineated, which in turn,

improves the accuracy of iterative PSI-BLAST searches. Notably, this improvement

to PSI-BLAST comes at minimal computational cost as SIB-BLAST utilizes existing

results already produced in a PSI-BLAST search.

DOI: 10.1093/nar/gkp301

PMCID: PMC2703926

PMID: 19429693 [Indexed for MEDLINE]

2213. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W101-5. doi: 10.1093/nar/gkp327.

Epub 2009 May 8.

Orphelia: predicting genes in metagenomic sequencing reads.

Hoff KJ(1), Lingner T, Meinicke P, Tech M.

Author information:

(1)Abteilung Bioinformatik, Institut für Mikrobiologie und Genetik,

Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany.

katharina@gobics.de

Metagenomic sequencing projects yield numerous sequencing reads of a diverse

range of uncultivated and mostly yet unknown microorganisms. In many cases, these

sequencing reads cannot be assembled into longer contigs. Thus, gene prediction

tools that were originally developed for whole-genome analysis are not suitable

for processing metagenomes. Orphelia is a program for predicting genes in short

DNA sequences that is available through a web server application

(http://orphelia.gobics.de). Orphelia utilizes prediction models that were

created with machine learning techniques on the basis of a wide range of

annotated genomes. In contrast to other methods for metagenomic gene prediction,

Orphelia has fragment length-specific prediction models for the two most popular

sequencing techniques in metagenomics, chain termination sequencing and

pyrosequencing. These models ensure highly specific gene predictions.

DOI: 10.1093/nar/gkp327

PMCID: PMC2703946

PMID: 19429689 [Indexed for MEDLINE]

2214. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W504-9. doi: 10.1093/nar/gkp324.

Epub 2009 May 8.

wwLigCSRre: a 3D ligand-based server for hit identification and optimization.

Sperandio O(1), Petitjean M, Tuffery P.

Author information:

(1)MTi, INSERM UMR-S973, Université Paris Diderot - Paris 7, F75013, Paris,

France.

The wwLigCSRre web server performs ligand-based screening using a 3D molecular

similarity engine. Its aim is to provide an online versatile facility to assist

the exploration of the chemical similarity of families of compounds, or to

propose some scaffold hopping from a query compound. The service allows the user

to screen several chemically diversified focused banks, such as Kinase-, CNS-,

GPCR-, Ion-channel-, Antibacterial-, Anticancer- and Analgesic-focused libraries.

The server also provides the possibility to screen the DrugBank and

DSSTOX/Carcinogenic compounds databases. User banks can also been downloaded. The

3D similarity search combines both geometrical (3D) and physicochemical

information. Starting from one 3D ligand molecule as query, the screening of such

databases can lead to unraveled compound scaffold as hits or help to optimize

previously identified hit molecules in a SAR (Structure activity relationship)

project. wwLigCSRre can be accessed at

http://bioserv.rpbs.univ-paris-diderot.fr/wwLigCSRre.html.

DOI: 10.1093/nar/gkp324

PMCID: PMC2703967

PMID: 19429687 [Indexed for MEDLINE]

2215. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W643-6. doi: 10.1093/nar/gkp321.

Epub 2009 May 8.

ATIVS: analytical tool for influenza virus surveillance.

Liao YC(1), Ko CY, Tsai MH, Lee MS, Hsiung CA.

Author information:

(1)Division of Biostatistics and Bioinformatics, Institute of Population Health

Sciences, National Health Research Institutes, Zhunan 350, Miaoli County, Taiwan.

The WHO Global Influenza Surveillance Network has routinely performed genetic and

antigenic analyses of human influenza viruses to monitor influenza activity.

Although these analyses provide supporting data for the selection of vaccine

strains, it seems desirable to have user-friendly tools to visualize the

antigenic evolution of influenza viruses for the purpose of surveillance. To meet

this need, we have developed a web server, ATIVS (Analytical Tool for Influenza

Virus Surveillance), for analyzing serological data of all influenza viruses and

hemagglutinin sequence data of human influenza A/H3N2 viruses so as to generate

antigenic maps for influenza surveillance and vaccine strain selection.

Functionalities are described and examples are provided to illustrate its

usefulness and performance. The ATIVS web server is available at

http://influenza.nhri.org.tw/ATIVS/.

DOI: 10.1093/nar/gkp321

PMCID: PMC2703974

PMID: 19429686 [Indexed for MEDLINE]

2216. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W510-4. doi: 10.1093/nar/gkp322.

Epub 2009 May 8.

QMEAN server for protein model quality estimation.

Benkert P(1), Künzli M, Schwede T.

Author information:

(1)Biozentrum, University of Basel, Switzerland.

Model quality estimation is an essential component of protein structure

prediction, since ultimately the accuracy of a model determines its usefulness

for specific applications. Usually, in the course of protein structure prediction

a set of alternative models is produced, from which subsequently the most

accurate model has to be selected. The QMEAN server provides access to two

scoring functions successfully tested at the eighth round of the community-wide

blind test experiment CASP. The user can choose between the composite scoring

function QMEAN, which derives a quality estimate on the basis of the geometrical

analysis of single models, and the clustering-based scoring function QMEANclust

which calculates a global and local quality estimate based on a weighted

all-against-all comparison of the models from the ensemble provided by the user.

The web server performs a ranking of the input models and highlights potentially

problematic regions for each model. The QMEAN server is available at

http://swissmodel.expasy.org/qmean.

DOI: 10.1093/nar/gkp322

PMCID: PMC2703985

PMID: 19429685 [Indexed for MEDLINE]

2217. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W422-7. doi: 10.1093/nar/gkp336.

Epub 2009 May 6.

PHEMTO: protein pH-dependent electric moment tools.

Kantardjiev AA(1), Atanasov BP.

Author information:

(1)Biophysical Chemistry Group, Institute of Organic Chemistry, Bulgarian Academy

of Sciences, Sofia-1113, Bulgaria.

PHEMTO (protein pH-dependent electric moment tools) is released in response to

the high demand in protein science community for evaluation of electrostatic

characteristics in relations to molecular recognition. PHEMTO will serve protein

scientists with new advanced features for analysis of protein molecular

interactions: Electric/dipole moments, their pH-dependence and in silico charge

mutagenesis effects on these properties as well as alternative algorithms for

electric/dipole moment computation--Singular value decomposition of electrostatic

potential (EP) to account for reaction field. The implementation is based on

long-term experience--PHEI mean field electrostatics and PHEPS server for

evaluation of global and local pH-dependent properties. However, PHEMTO is not

just an update of our PHEPS server. Besides standard electrostatics, we offer

new, advanced and useful features for analysis of protein molecular interactions.

In addition our algorithms are very fast. Special emphasis is given to the

interface--intuitive and user-friendly. The input is comprised of the atomic

coordinate file in Protein Data Bank format. The advanced user is provided with a

special input section for addition of non-polypeptide charges. The output covers

actually full electrostatic characteristics but special emphasis is given to

electric/dipole moments and their interactive visualization. PHEMTO server can be

accessed at http://phemto.orgchm.bas.bg/.

DOI: 10.1093/nar/gkp336

PMCID: PMC2703894

PMID: 19420068 [Indexed for MEDLINE]

2218. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W617-22. doi:

10.1093/nar/gkp293. Epub 2009 May 6.

OptiTope--a web server for the selection of an optimal set of peptides for

epitope-based vaccines.

Toussaint NC(1), Kohlbacher O.

Author information:

(1)Division for Simulation of Biological Systems, Wilhelm Schickard Institute for

Computer Science, Center for Bioinformatics Tübingen, Eberhard-Karls-Universität

Tübingen, Germany. toussaint@informatik.uni-tuebingen.de

Epitope-based vaccines (EVs) have recently been attracting significant interest.

They trigger an immune response by confronting the immune system with immunogenic

peptides derived from, e.g. viral- or cancer-related proteins. Binding of these

peptides to proteins from the major histocompatibility complex (MHC) is crucial

for immune system activation. However, since the MHC is highly polymorphic,

different patients typically bind different repertoires of peptides. Furthermore,

economical and regulatory issues impose strong limitations on the number of

peptides that can be included in an EV. Hence, it is crucial to identify the

optimal set of peptides for a vaccine, given constraints such as MHC allele

probabilities in the target population, peptide mutation rates and maximum number

of selected peptides. OptiTope aims at assisting immunologists in this critical

task. With OptiTope, we provide an easy-to-use tool to determine a provably

optimal set of epitopes with respect to overall immunogenicity in a specific

individual (personalized medicine) or a target population (e.g. a certain ethnic

group). OptiTope is available at http://www.epitoolkit.org/optitope.

DOI: 10.1093/nar/gkp293

PMCID: PMC2703925

PMID: 19420066 [Indexed for MEDLINE]

2219. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W48-52. doi: 10.1093/nar/gkp279.

Epub 2009 May 6.

webPRC: the Profile Comparer for alignment-based searching of public domain

databases.

Brandt BW(1), Heringa J.

Author information:

(1)Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam, The

Netherlands. bwbrandt@few.vu.nl

Profile-profile methods are well suited to detect remote evolutionary

relationships between protein families. Profile Comparer (PRC) is an existing

stand-alone program for scoring and aligning hidden Markov models (HMMs), which

are based on multiple sequence alignments. Since PRC compares profile HMMs

instead of sequences, it can be used to find distant homologues. For this

purpose, PRC is used by, for example, the CATH and Pfam-domain databases. As PRC

is a profile comparer, it only reports profile HMM alignments and does not

produce multiple sequence alignments. We have developed webPRC server, which

makes it straightforward to search for distant homologues or similar alignments

in a number of domain databases. In addition, it provides the results both as

multiple sequence alignments and aligned HMMs. Furthermore, the user can view the

domain annotation, evaluate the PRC hits with the Jalview multiple alignment

editor and generate logos from the aligned HMMs or the aligned multiple

alignments. Thus, this server assists in detecting distant homologues with PRC as

well as in evaluating and using the results. The webPRC interface is available at

http://www.ibi.vu.nl/programs/prcwww/.

DOI: 10.1093/nar/gkp279

PMCID: PMC2703954

PMID: 19420063 [Indexed for MEDLINE]

2220. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W515-8. doi: 10.1093/nar/gkp305.

Epub 2009 May 6.

NNcon: improved protein contact map prediction using 2D-recursive neural

networks.

Tegge AN(1), Wang Z, Eickholt J, Cheng J.

Author information:

(1)Computer Science Department, Informatics Institute, University of Missouri,

Columbia, MO 65213, USA.

Protein contact map prediction is useful for protein folding rate prediction,

model selection and 3D structure prediction. Here we describe NNcon, a fast and

reliable contact map prediction server and software. NNcon was ranked among the

most accurate residue contact predictors in the Eighth Critical Assessment of

Techniques for Protein Structure Prediction (CASP8), 2008. Both NNcon server and

software are available at http://casp.rnet.missouri.edu/nncon.html.

DOI: 10.1093/nar/gkp305

PMCID: PMC2703959

PMID: 19420062 [Indexed for MEDLINE]

2221. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W526-31. doi:

10.1093/nar/gkp316. Epub 2009 May 6.

ProSMoS server: a pattern-based search using interaction matrix representation of

protein structures.

Shi S(1), Chitturi B, Grishin NV.

Author information:

(1)Howard Hughes Medical Institute, University of Texas Southwestern Medical

Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA.

Assessing structural similarity and defining common regions through comparison of

protein spatial structures is an important task in functional and evolutionary

studies of proteins. There are many servers that compare structures and define

sub-structures in common between proteins through superposition and closeness of

either coordinates or contacts. However, a natural way to analyze a structure for

experts working on structure classification is to look for specific

three-dimensional (3D) motifs and patterns instead of finding common features in

two proteins. Such motifs can be described by the architecture and topology of

major secondary structural elements (SSEs) without consideration of subtle

differences in 3D coordinates. Despite the importance of motif-based structure

searches, currently there is a shortage of servers to perform this task. Widely

known TOPS does not fully address this problem, as it finds only topological

match but does not take into account other important spatial properties, such as

interactions and chirality. Here, we implemented our approach to protein

structure pattern search (ProSMoS) as a web-server. ProSMoS converts 3D structure

into an interaction matrix representation including the SSE types, handednesses

of connections between SSEs, coordinates of SSE starts and ends, types of

interactions between SSEs and beta-sheet definitions. For a user-defined

structure pattern, ProSMoS lists all structures from a database that contain this

pattern. ProSMoS server will be of interest to structural biologists who would

like to analyze very general and distant structural similarities. The ProSMoS web

server is available at: http://prodata.swmed.edu/ProSMoS/.

DOI: 10.1093/nar/gkp316

PMCID: PMC2703969

PMID: 19420061 [Indexed for MEDLINE]

2222. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W545-51. doi:

10.1093/nar/gkp291. Epub 2009 May 6.

iSARST: an integrated SARST web server for rapid protein structural similarity

searches.

Lo WC(1), Lee CY, Lee CC, Lyu PC.

Author information:

(1)Institute of Bioinformatics and Structural Biology, National Tsing Hua

University, Hsinchu, Taiwan.

iSARST is a web server for efficient protein structural similarity searches. It

is a multi-processor, batch-processing and integrated implementation of several

structural comparison tools and two database searching methods: SARST for common

structural homologs and CPSARST for homologs with circular permutations. iSARST

allows users submitting multiple PDB/SCOP entry IDs or an archive file containing

many structures. After scanning the target database using SARST/CPSARST, the

ordering of hits are refined with conventional structure alignment tools such as

FAST, TM-align and SAMO, which are run in a PC cluster. In this way, iSARST

achieves a high running speed while preserving the high precision of refinement

engines. The final outputs include tables listing co-linear or circularly

permuted homologs of the query proteins and a functional summary of the best

hits. Superimposed structures can be examined through an interactive and

informative visualization tool. iSARST provides the first batch mode structural

comparison web service for both co-linear homologs and circular permutants. It

can serve as a rapid annotation system for functionally unknown or hypothetical

proteins, which are increasing rapidly in this post-genomics era. The server can

be accessed at http://sarst.life.nthu.edu.tw/iSARST/.

DOI: 10.1093/nar/gkp291

PMCID: PMC2703971

PMID: 19420060 [Indexed for MEDLINE]

2223. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W519-25. doi:

10.1093/nar/gkp306. Epub 2009 May 6.

Protinfo PPC: a web server for atomic level prediction of protein complexes.

Kittichotirat W(1), Guerquin M, Bumgarner RE, Samudrala R.

Author information:

(1)Department of Microbiology, University of Washington, Seattle, WA, USA.

'Protinfo PPC' (Prediction of Protein Complex) is a web server that predicts

atomic level structures of interacting proteins from their amino-acid sequences.

It uses the interolog method to search for experimental protein complex

structures that are homologous to the input sequences submitted by a user. These

structures are then used as starting templates to generate protein complex

models, which are returned to the user in Protein Data Bank format via email. The

server supports modeling of both homo and hetero multimers and generally produces

full atomic level models (including insertion/deletion regions) of protein

complexes as long as at least one putative homologous template for the query

sequences is found. The modeling pipeline behind Protinfo PPC has been rigorously

benchmarked and proven to produce highly accurate protein complex models. The

fully automated all atom comparative modeling service for protein complexes

provided by Protinfo PPC server offers wide capabilities ranging from prediction

of protein complex interactions to identification of possible interaction sites,

which will be useful for researchers studying these topics. The Protinfo PPC web

server is available at http://protinfo.compbio.washington.edu/ppc/.

DOI: 10.1093/nar/gkp306

PMCID: PMC2703994

PMID: 19420059 [Indexed for MEDLINE]

2224. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W23-7. doi: 10.1093/nar/gkp265.

Epub 2009 May 6.

BioMart Central Portal--unified access to biological data.

Haider S(1), Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A.

Author information:

(1)EMBL-European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, Computer

Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK.

BioMart Central Portal (www.biomart.org) offers a one-stop shop solution to

access a wide array of biological databases. These include major biomolecular

sequence, pathway and annotation databases such as Ensembl, Uniprot, Reactome,

HGNC, Wormbase and PRIDE; for a complete list, visit,

http://www.biomart.org/biomart/martview. Moreover, the web server features

seamless data federation making cross querying of these data sources in a user

friendly and unified way. The web server not only provides access through a web

interface (MartView), it also supports programmatic access through a Perl API as

well as RESTful and SOAP oriented web services. The website is free and open to

all users and there is no login requirement.

DOI: 10.1093/nar/gkp265

PMCID: PMC2703988

PMID: 19420058 [Indexed for MEDLINE]

2225. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W532-8. doi: 10.1093/nar/gkp328.

Epub 2009 May 5.

HorA web server to infer homology between proteins using sequence and structural

similarity.

Kim BH(1), Cheng H, Grishin NV.

Author information:

(1)Department of Biochemistry, University of Texas, Southwestern Medical Center,

5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA.

The biological properties of proteins are often gleaned through comparative

analysis of evolutionary relatives. Although protein structure similarity search

methods detect more distant homologs than purely sequence-based methods,

structural resemblance can result from either homology (common ancestry) or

analogy (similarity without common ancestry). While many existing web servers

detect structural neighbors, they do not explicitly address the question of

homology versus analogy. Here, we present a web server named HorA (Homology or

Analogy) that identifies likely homologs for a query protein structure. Unlike

other servers, HorA combines sequence information from state-of-the-art profile

methods with structure information from spatial similarity measures using an

advanced computational technique. HorA aims to identify biologically meaningful

connections rather than purely 3D-geometric similarities. The HorA method finds

approximately 90% of remote homologs defined in the manually curated database

SCOP. HorA will be especially useful for finding remote homologs that might be

overlooked by other sequence or structural similarity search servers. The HorA

server is available at http://prodata.swmed.edu/horaserver.

DOI: 10.1093/nar/gkp328

PMCID: PMC2703895

PMID: 19417074 [Indexed for MEDLINE]

2226. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W235-9. doi: 10.1093/nar/gkp287.

Epub 2009 May 5.

3D-DART: a DNA structure modelling server.

van Dijk M(1), Bonvin AM.

Author information:

(1)Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University,

The Netherlands.

There is a growing interest in structural studies of DNA by both experimental and

computational approaches. Often, 3D-structural models of DNA are required, for

instance, to serve as templates for homology modeling, as starting structures for

macro-molecular docking or as scaffold for NMR structure calculations. The

conformational adaptability of DNA when binding to a protein is often an

important factor and at the same time a limitation in such studies. As a response

to the demand for 3D-structural models reflecting the intrinsic plasticity of DNA

we present the 3D-DART server (3DNA-Driven DNA Analysis and Rebuilding Tool). The

server provides an easy interface to a powerful collection of tools for the

generation of DNA-structural models in custom conformations. The computational

engine beyond the server makes use of the 3DNA software suite together with a

collection of home-written python scripts. The server is freely available at

http://haddock.chem.uu.nl/dna without any login requirement.

DOI: 10.1093/nar/gkp287

PMCID: PMC2703913

PMID: 19417072 [Indexed for MEDLINE]

2227. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W229-34. doi:

10.1093/nar/gkp286. Epub 2009 May 5.

Insignia: a DNA signature search web server for diagnostic assay development.

Phillippy AM(1), Ayanbule K, Edwards NJ, Salzberg SL.

Author information:

(1)Center for Bioinformatics and Computational Biology, University of Maryland,

College Park, MD, USA. amp@umiacs.umd.edu

Insignia is a web application for the rapid identification of unique DNA

signatures. DNA signatures are distinct nucleotide sequences that can be used to

detect the presence of certain organisms and to distinguish those organisms from

all other species. These signatures can be used as the basis for diagnostic

assays to detect and genotype microbes in both environmental and clinical

samples. Insignia identifies an exhaustive set of accurate DNA signatures for any

set of target genomes, and screens these signatures against a comprehensive

background that includes all sequenced bacteria and viruses, the human genome,

and many other animals and plants. Identified signatures may be browsed by

genomic location or proximal genes, filtered by composition, viewed in a genome

browser or directly downloaded. Integrated PCR primer design is also provided for

each signature. The Insignia website (http://insignia.cbcb.umd.edu) is free and

open to all users and there is no login requirement. In addition, the source code

for the computational pipeline is freely available.

DOI: 10.1093/nar/gkp286

PMCID: PMC2703920

PMID: 19417071 [Indexed for MEDLINE]

2228. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W369-75. doi:

10.1093/nar/gkp309. Epub 2009 May 5.

PPISearch: a web server for searching homologous protein-protein interactions

across multiple species.

Chen CC(1), Lin CY, Lo YS, Yang JM.

Author information:

(1)Institute of Bioinformatics and Systems Biology, National Chiao Tung

University, Hsinchu, 30050, Taiwan.

As an increasing number of reliable protein-protein interactions (PPIs) become

available and high-throughput experimental methods provide systematic

identification of PPIs, there is a growing need for fast and accurate methods for

discovering homologous PPIs of a newly determined PPI. PPISearch is a web server

that rapidly identifies homologous PPIs (called PPI family) and infers

transferability of interacting domains and functions of a query protein pair.

This server first identifies two homologous families of the query, respectively,

by using BLASTP to scan an annotated PPIs database (290 137 PPIs in 576 species),

which is a collection of five public databases. We determined homologous PPIs

from protein pairs of homologous families when these protein pairs were in the

annotated database and have significant joint sequence similarity (E < or =

10(-40)) with the query. Using these homologous PPIs across multiple species,

this sever infers the conserved domain-domain pairs (Pfam and InterPro domains)

and function pairs (Gene Ontology annotations). Our results demonstrate that the

transferability of conserved domain-domain pairs between homologous PPIs and

query pairs is 88% using 103 762 PPI queries, and the transferability of

conserved function pairs is 69% based on 106 997 PPI queries. The PPISearch

server should be useful for searching homologous PPIs and PPI families across

multiple species. The PPISearch server is available through the website at

http://gemdock.life.nctu.edu.tw/ppisearch/.

DOI: 10.1093/nar/gkp309

PMCID: PMC2703927

PMID: 19417070 [Indexed for MEDLINE]

2229. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W406-12. doi:

10.1093/nar/gkp312. Epub 2009 May 5.

SePreSA: a server for the prediction of populations susceptible to serious

adverse drug reactions implementing the methodology of a chemical-protein

interactome.

Yang L(1), Luo H, Chen J, Xing Q, He L.

Author information:

(1)Bio-X Center, Key Laboratory for the Genetics of Developmental and

Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong

University, Shanghai 200030, PR China. lunyang@gmail.com

Serious adverse drug reactions (SADRs) are caused by unexpected drug-human

protein interactions, and some polymorphisms within binding pockets make the

population carrying these polymorphisms susceptible to SADR. Predicting which

populations are likely to be susceptible to SADR will not only strengthen drug

safety, but will also assist enterprises to adjust R&D and marketing strategies.

Making such predictions has recently been facilitated by the introduction of a

web server named SePreSA. The server has a comprehensive collection of the

structural models of nearly all the well known SADR targets. Once a drug molecule

is submitted, the scale of its potential interaction with multi-SADR targets is

calculated using the DOCK program. The server utilizes a 2-directional

Z-transformation scoring algorithm, which computes the relative drug-protein

interaction strength based on the docking-score matrix of a chemical-protein

interactome, thus achieve greater accuracy in prioritizing SADR targets than

simply using dock scoring functions. The server also suggests the binding pattern

of the lowest docking score through 3D visualization, by highlighting and

visualizing amino acid residues involved in the binding on the customer's

browser. Polymorphism information for different populations for each of the

interactive residues will be displayed, helping users to deduce the

population-specific susceptibility of their drug molecule. The server is freely

available at http://SePreSA.Bio-X.cn/.

DOI: 10.1093/nar/gkp312

PMCID: PMC2703957

PMID: 19417066 [Indexed for MEDLINE]

2230. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W296-9. doi: 10.1093/nar/gkp268.

Epub 2009 Apr 30.

The Microbe browser for comparative genomics.

Gattiker A(1), Dessimoz C, Schneider A, Xenarios I, Pagni M, Rougemont J.

Author information:

(1)Swiss Institute of Bioinformatics, School of Life Sciences, Ecole

Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland.

The Microbe browser is a web server providing comparative microbial genomics

data. It offers comprehensive, integrated data from GenBank, RefSeq, UniProt,

InterPro, Gene Ontology and the Orthologs Matrix Project (OMA) database,

displayed along with gene predictions from five software packages. The Microbe

browser is daily updated from the source databases and includes all completely

sequenced bacterial and archaeal genomes. The data are displayed in an

easy-to-use, interactive website based on Ensembl software. The Microbe browser

is available at http://microbe.vital-it.ch/. Programmatic access is available

through the OMA application programming interface (API) at

http://microbe.vital-it.ch/api.

DOI: 10.1093/nar/gkp268

PMCID: PMC2703916

PMID: 19406928 [Indexed for MEDLINE]

2231. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W670-7. doi: 10.1093/nar/gkp280.

Epub 2009 Apr 30.

GeNMR: a web server for rapid NMR-based protein structure determination.

Berjanskii M(1), Tang P, Liang J, Cruz JA, Zhou J, Zhou Y, Bassett E, MacDonell

C, Lu P, Lin G, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta and National Research

Council, National Institute for Nanotechnology, Edmonton, AB, Canada T6G 2E8.

GeNMR (GEnerate NMR structures) is a web server for rapidly generating accurate

3D protein structures using sequence data, NOE-based distance restraints and/or

NMR chemical shifts as input. GeNMR accepts distance restraints in XPLOR or CYANA

format as well as chemical shift files in either SHIFTY or BMRB formats. The web

server produces an ensemble of PDB coordinates for the protein within 15-25 min,

depending on model complexity and completeness of experimental restraints. GeNMR

uses a pipeline of several pre-existing programs and servers to calculate the

actual protein structure. In particular, GeNMR combines genetic algorithms for

structure optimization along with homology modeling, chemical shift threading,

torsion angle and distance predictions from chemical shifts/NOEs as well as

ROSETTA-based structure generation and simulated annealing with XPLOR-NIH to

generate and/or refine protein coordinates. GeNMR greatly simplifies the task of

protein structure determination as users do not have to install or become

familiar with complex stand-alone programs or obscure format conversion

utilities. Tests conducted on a sample of 90 proteins from the BioMagResBank

indicate that GeNMR produces high-quality models for all protein queries,

regardless of the type of NMR input data. GeNMR was developed to facilitate

rapid, user-friendly structure determination of protein structures via NMR

spectroscopy. GeNMR is accessible at http://www.genmr.ca.

DOI: 10.1093/nar/gkp280

PMCID: PMC2703936

PMID: 19406927 [Indexed for MEDLINE]

2232. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W329-34. doi:

10.1093/nar/gkp263. Epub 2009 Apr 30.

WhichGenes: a web-based tool for gathering, building, storing and exporting gene

sets with application in gene set enrichment analysis.

Glez-Peña D(1), Gómez-López G, Pisano DG, Fdez-Riverola F.

Author information:

(1)Higher Technical School of Computer Engineering, University of Vigo, Ourense,

Spain.

WhichGenes is a web-based interactive gene set building tool offering a very

simple interface to extract always-updated gene lists from multiple databases and

unstructured biological data sources. While the user can specify new gene sets of

interest by following a simple four-step wizard, the tool is able to run several

queries in parallel. Every time a new set is generated, it is automatically added

to the private gene-set cart and the user is notified by an e-mail containing a

direct link to the new set stored in the server. WhichGenes provides

functionalities to edit, delete and rename existing sets as well as the

capability of generating new ones by combining previous existing sets

(intersection, union and difference operators). The user can export his sets

configuring the output format and selecting among multiple gene identifiers. In

addition to the user-friendly environment, WhichGenes allows programmers to

access its functionalities in a programmatic way through a Representational State

Transfer web service. WhichGenes front-end is freely available at

http://www.whichgenes.org/, WhichGenes API is accessible at

http://www.whichgenes.org/api/.

DOI: 10.1093/nar/gkp263

PMCID: PMC2703947

PMID: 19406925 [Indexed for MEDLINE]

2233. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W273-6. doi: 10.1093/nar/gkp292.

Epub 2009 Apr 30.

DIANA-microT web server: elucidating microRNA functions through target

prediction.

Maragkakis M(1), Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T,

Giannopoulos G, Goumas G, Koukis E, Kourtis K, Vergoulis T, Koziris N, Sellis T,

Tsanakas P, Hatzigeorgiou AG.

Author information:

(1)Department of Molecular Oncology, Biomedical Sciences Research Center

Alexander Fleming, Vari.

Computational microRNA (miRNA) target prediction is one of the key means for

deciphering the role of miRNAs in development and disease. Here, we present the

DIANA-microT web server as the user interface to the DIANA-microT 3.0 miRNA

target prediction algorithm. The web server provides extensive information for

predicted miRNA:target gene interactions with a user-friendly interface,

providing extensive connectivity to online biological resources. Target gene and

miRNA functions may be elucidated through automated bibliographic searches and

functional information is accessible through Kyoto Encyclopedia of Genes and

Genomes (KEGG) pathways. The web server offers links to nomenclature, sequence

and protein databases, and users are facilitated by being able to search for

targeted genes using different nomenclatures or functional features, such as the

genes possible involvement in biological pathways. The target prediction

algorithm supports parameters calculated individually for each miRNA:target gene

interaction and provides a signal-to-noise ratio and a precision score that helps

in the evaluation of the significance of the predicted results. Using a set of

miRNA targets recently identified through the pSILAC method, the performance of

several computational target prediction programs was assessed. DIANA-microT 3.0

achieved there with 66% the highest ratio of correctly predicted targets over all

predicted targets. The DIANA-microT web server is freely available at

www.microrna.gr/microT.

DOI: 10.1093/nar/gkp292

PMCID: PMC2703977

PMID: 19406924 [Indexed for MEDLINE]

2234. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W384-9. doi: 10.1093/nar/gkp308.

Epub 2009 Apr 30.

SplitPocket: identification of protein functional surfaces and characterization

of their spatial patterns.

Tseng YY(1), Dupree C, Chen ZJ, Li WH.

Author information:

(1)Department of Ecology and Evolution, University of Chicago, 1101 East 57th

Street, Chicago, IL 60637, USA.

SplitPocket (http://pocket.uchicago.edu/) is a web server to identify functional

surfaces of protein from structure coordinates. Using the Alpha Shape Theory, we

previously developed an analytical approach to identify protein functional

surfaces by the geometric concept of a split pocket, which is a pocket split by a

binding ligand. Our geometric approach extracts site-specific spatial information

from coordinates of structures. To reduce the search space, probe radii are

designed according to the physicochemical textures of molecules. The method uses

the weighted Delaunay triangulation and the discrete flow algorithm to obtain

geometric measurements and spatial patterns for each predicted pocket. It can

also measure the hydrophobicity on a surface patch. Furthermore, we quantify the

evolutionary conservation of surface patches by an index derived from the entropy

scores in HSSP (homology-derived secondary structure of proteins). We have used

the method to examine approximately 1.16 million potential pockets and identified

the split pockets in >26,000 structures in the Protein Data Bank. This integrated

web server of functional surfaces provides a source of spatial patterns to serve

as templates for predicting the functional surfaces of unbound structures

involved in binding activities. These spatial patterns should also be useful for

protein functional inference, structural evolution and drug design.

DOI: 10.1093/nar/gkp308

PMCID: PMC2703984

PMID: 19406922 [Indexed for MEDLINE]

2235. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W129-34. doi:

10.1093/nar/gkp264. Epub 2009 Apr 28.

FMM: a web server for metabolic pathway reconstruction and comparative analysis.

Chou CH(1), Chang WC, Chiu CM, Huang CC, Huang HD.

Author information:

(1)Institute of Molecular Medicine and Bioengineering, National Chiao Tung

University, Hsin-Chu 300, Taiwan, Republic of China.

Synthetic Biology, a multidisciplinary field, is growing rapidly. Improving the

understanding of biological systems through mimicry and producing bio-orthogonal

systems with new functions are two complementary pursuits in this field. A web

server called FMM (From Metabolite to Metabolite) was developed for this purpose.

FMM can reconstruct metabolic pathways form one metabolite to another metabolite

among different species, based mainly on the Kyoto Encyclopedia of Genes and

Genomes (KEGG) database and other integrated biological databases. Novel

presentation for connecting different KEGG maps is newly provided. Both local and

global graphical views of the metabolic pathways are designed. FMM has many

applications in Synthetic Biology and Metabolic Engineering. For example, the

reconstruction of metabolic pathways to produce valuable metabolites or secondary

metabolites in bacteria or yeast is a promising strategy for drug production. FMM

provides a highly effective way to elucidate the genes from which species should

be cloned into those microorganisms based on FMM pathway comparative analysis.

Consequently, FMM is an effective tool for applications in synthetic biology to

produce both drugs and biofuels. This novel and innovative resource is now freely

available at http://FMM.mbc.nctu.edu.tw/.

DOI: 10.1093/nar/gkp264

PMCID: PMC2703958

PMID: 19401437 [Indexed for MEDLINE]

2236. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W209-13. doi:

10.1093/nar/gkp269. Epub 2009 Apr 28.

UniPrime2: a web service providing easier Universal Primer design.

Boutros R(1), Stokes N, Bekaert M, Teeling EC.

Author information:

(1)UCD School of Computer Science and Informatics, University College Dublin,

Belfield, Dublin 4, Ireland.

The UniPrime2 web server is a publicly available online resource which

automatically designs large sets of universal primers when given a gene reference

ID or Fasta sequence input by a user. UniPrime2 works by automatically retrieving

and aligning homologous sequences from GenBank, identifying regions of

conservation within the alignment, and generating suitable primers that can be

used to amplify variable genomic regions. In essence, UniPrime2 is a suite of

publicly available software packages (Blastn, T-Coffee, GramAlign, Primer3),

which reduces the laborious process of primer design, by integrating these

programs into a single software pipeline. Hence, UniPrime2 differs from previous

primer design web services in that all steps are automated, linked, saved and

phylogenetically delimited, only requiring a single user-defined gene reference

ID or input sequence. We provide an overview of the web service and

wet-laboratory validation of the primers generated. The system is freely

accessible at: http://uniprime.batlab.eu. UniPrime2 is licenced under a Creative

Commons Attribution Noncommercial-Share Alike 3.0 Licence.

DOI: 10.1093/nar/gkp269

PMCID: PMC2703989

PMID: 19401435 [Indexed for MEDLINE]

2237. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W413-6. doi: 10.1093/nar/gkp281.

Epub 2009 Apr 26.

SITEHOUND-web: a server for ligand binding site identification in protein

structures.

Hernandez M(1), Ghersi D, Sanchez R.

Author information:

(1)Department of Structural and Chemical Biology, Mount Sinai School of Medicine,

New York, NY 10029, USA.

SITEHOUND-web (http://sitehound.sanchezlab.org) is a binding-site identification

server powered by the SITEHOUND program. Given a protein structure in PDB format

SITEHOUND-web will identify regions of the protein characterized by favorable

interactions with a probe molecule. These regions correspond to putative ligand

binding sites. Depending on the probe used in the calculation, sites with

preference for different ligands will be identified. Currently, a carbon probe

for identification of binding sites for drug-like molecules, and a phosphate

probe for phosphorylated ligands (ATP, phoshopeptides, etc.) have been

implemented. SITEHOUND-web will display the results in HTML pages including an

interactive 3D representation of the protein structure and the putative sites

using the Jmol java applet. Various downloadable data files are also provided for

offline data analysis.

DOI: 10.1093/nar/gkp281

PMCID: PMC2703923

PMID: 19398430 [Indexed for MEDLINE]

2238. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W174-8. doi: 10.1093/nar/gkp278.

Epub 2009 Apr 26.

CVTree update: a newly designed phylogenetic study platform using composition

vectors and whole genomes.

Xu Z(1), Hao B.

Author information:

(1)T-Life Research Center, Fudan University, 220 Handan Road, Shanghai 200433,

China. xuzh.fdu@gmail.com

The CVTree web server (http://tlife.fudan.edu.cn/cvtree) presented here is a new

implementation of the whole genome-based, alignment-free composition vector (CV)

method for phylogenetic analysis. It is more efficient and user-friendly than the

previously published version in the 2004 web server issue of Nucleic Acids

Research. The development of whole genome-based alignment-free CV method has

provided an independent verification to the traditional phylogenetic analysis

based on a single gene or a few genes. This new implementation attempts to meet

the challenge of ever increasing amount of genome data and includes in its

database more than 850 prokaryotic genomes which will be updated monthly from

NCBI, and more than 80 fungal genomes collected manually from several sequencing

centers. This new CVTree web server provides a faster and stable research

platform. Users can upload their own sequences to find their phylogenetic

position among genomes selected from the server's; inbuilt database. All sequence

data used in a session may be downloaded as a compressed file. In addition to

standard phylogenetic trees, users can also choose to output trees whose

monophyletic branches are collapsed to various taxonomic levels. This feature is

particularly useful for comparing phylogeny with taxonomy when dealing with

thousands of genomes.

DOI: 10.1093/nar/gkp278

PMCID: PMC2703908

PMID: 19398429 [Indexed for MEDLINE]

2239. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W402-5. doi: 10.1093/nar/gkp256.

Epub 2009 Apr 24.

ProteinCCD: enabling the design of protein truncation constructs for expression

and crystallization experiments.

Mooij WT(1), Mitsiki E, Perrakis A.

Author information:

(1)Department of Biochemistry, NKI, Plesmanlaan 121, 1066 CX, Amsterdam, The

Netherlands.

ProteinCCD (CCD for Crystallographic Construct Design) aims to facilitate a

common practice in structural biology, namely the design of several truncation

constructs of the protein under investigation, based on experimental data or on

sequence analysis tools. ProteinCCD functions as a meta-server, available online

at http://xtal.nki.nl/ccd, that collects information from prediction servers

concerning secondary structure, disorder, coiled coils, transmembrane segments,

domains and domain linkers. It then displays a condensed view of all results

against the protein sequence. The user can study the output and choose

interactively possible starts and ends for suitable protein constructs. Since the

required input to ProteinCCD is the DNA and not the protein sequence, once the

starts and ends of constructs are chosen, the software can automatically design

the oligonucleotides needed for PCR amplification of all constructs. ProteinCCD

outputs a comprehensive view of all constructs and all oligos needed for

bookkeeping or for direct copy-paste ordering of the designed oligonucleotides.

DOI: 10.1093/nar/gkp256

PMCID: PMC2703965

PMID: 19395596 [Indexed for MEDLINE]

2240. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W441-5. doi: 10.1093/nar/gkp253.

Epub 2009 Apr 24.

IC50-to-Ki: a web-based tool for converting IC50 to Ki values for inhibitors of

enzyme activity and ligand binding.

Cer RZ(1), Mudunuri U, Stephens R, Lebeda FJ.

Author information:

(1)Advanced Biomedical Computing Center, Advanced Technology Program,

SAIC-Frederick Inc., NCI-Frederick, Frederick, MD 21702, USA.

A new web-server tool estimates K(i) values from experimentally determined IC(50)

values for inhibitors of enzymes and of binding reactions between macromolecules

(e.g. proteins, polynucleic acids) and ligands. This converter was developed to

enable end users to help gauge the quality of the underlying assumptions used in

these calculations which depend on the type of mechanism of inhibitor action and

the concentrations of the interacting molecular species. Additional calculations

are performed for nonclassical, tightly bound inhibitors of enzyme-substrate or

of macromolecule-ligand systems in which free, rather than total concentrations

of the reacting species are required. Required user-defined input values include

the total enzyme (or another target molecule) and substrate (or ligand)

concentrations, the K(m) of the enzyme-substrate (or the K(d) of the

target-ligand) reaction, and the IC(50) value. Assumptions and caveats for these

calculations are discussed along with examples taken from the literature. The

host database for this converter contains kinetic constants and other data for

inhibitors of the proteolytic clostridial neurotoxins

(http://botdb.abcc.ncifcrf.gov/toxin/kiConverter.jsp).

DOI: 10.1093/nar/gkp253

PMCID: PMC2703898

PMID: 19395593 [Indexed for MEDLINE]

2241. Protein Eng Des Sel. 2009 Jul;22(7):441-4. doi: 10.1093/protein/gzp016. Epub 2009

Jun 2.

Prediction and classification of chemokines and their receptors.

Lata S(1), Raghava GP.

Author information:

(1)Bioinformatics Center, Institute of Microbial Technology, Chandigarh, India.

Chemokines are low molecular mass cytokine-like proteins that orchestrate myriads

of immune functions like leukocyte trafficking, T cell differentiation,

angiogenesis, hematopeosis and mast cell degranulation. Chemokines also play a

role as HIV-1 inhibitor and act as potent natural adjuvant in antitumor

immunotherapy. Receptors for these molecules are all seven-pass transmembrane

G-protein-coupled receptors that are intimately involved with chemokines in a

wide array of physiological and pathological conditions. These receptors also

have a major role as co-receptors for HIV-1 entry into target cells. Therefore,

chemokine receptors have proven to be excellent targets for small molecule in

pharmaceutical industry. The immense importance of chemokines and their receptors

motivated us to develop a support vector machine-based method ChemoPred to

predict this important class of proteins and further classify them into

subfamilies. ChemoPred is capable of predicting chemokines and chemokine

receptors with an accuracy of 95.08% and 92.19%, respectively. The overall

accuracy of classification of chemokines into three subfamilies was 96.00% and

that of chemokine receptors into three families was 92.87%. The server ChemoPred

is freely available at www.imtech.res.in/raghava/chemopred.

DOI: 10.1093/protein/gzp016

PMID: 19491216 [Indexed for MEDLINE]

2242. Proteins. 2009 Jul;76(1):115-28. doi: 10.1002/prot.22323.

A generalized knowledge-based discriminatory function for biomolecular

interactions.

Bernard B(1), Samudrala R.

Author information:

(1)Department of Bioengineering, University of Washington, Seattle, WA 98195,

USA.

Several novel and established knowledge-based discriminatory function

formulations and reference state derivations have been evaluated to identify

parameter sets capable of distinguishing native and near-native biomolecular

interactions from incorrect ones. We developed the r.m.r function, a novel atomic

level radial distribution function with mean reference state that averages over

all pairwise atom types from a reduced atom type composition, using

experimentally determined intermolecular complexes in the Cambridge Structural

Database (CSD) and the Protein Data Bank (PDB) as the information sources. We

demonstrate that r.m.r had the best discriminatory accuracy and power for

protein-small molecule and protein-DNA interactions, regardless of whether the

native complex was included or excluded, from the test set. The superior

performance of the r.m.r discriminatory function compared with seventeen

alternative functions evaluated on publicly available test sets for protein-small

molecule and protein-DNA interactions indicated that the function was not over

optimized through back testing on a single class of biomolecular interactions.

The initial success of the reduced composition and superior performance with the

CSD as the distribution set over the PDB implies that further improvements and

generality of the function are possible by deriving probabilities from subsets of

the CSD, using structures that consist of only the atom types to be considered

for given biomolecular interactions. The method is available as a web server

module at http://protinfo.compbio.washington.edu.

DOI: 10.1002/prot.22323

PMCID: PMC2891153

PMID: 19127590 [Indexed for MEDLINE]

2243. RNA. 2009 Jul;15(7):1426-30. doi: 10.1261/rna.1623809. Epub 2009 May 21.

Computational identification of riboswitches based on RNA conserved functional

sequences and conformations.

Chang TH(1), Huang HD, Wu LC, Yeh CT, Liu BJ, Horng JT.

Author information:

(1)Department of Computer Science and Information Engineering, National Central

University, Jhongli 320, Taiwan.

Riboswitches are cis-acting genetic regulatory elements within a specific mRNA

that can regulate both transcription and translation by interacting with their

corresponding metabolites. Recently, an increasing number of riboswitches have

been identified in different species and investigated for their roles in

regulatory functions. Both the sequence contexts and structural conformations are

important characteristics of riboswitches. None of the previously developed

tools, such as covariance models (CMs), Riboswitch finder, and RibEx, provide a

web server for efficiently searching homologous instances of known riboswitches

or considers two crucial characteristics of each riboswitch, such as the

structural conformations and sequence contexts of functional regions. Therefore,

we developed a systematic method for identifying 12 kinds of riboswitches. The

method is implemented and provided as a web server, RiboSW, to efficiently and

conveniently identify riboswitches within messenger RNA sequences. The predictive

accuracy of the proposed method is comparable with other previous tools. The

efficiency of the proposed method for identifying riboswitches was improved in

order to achieve a reasonable computational time required for the prediction,

which makes it possible to have an accurate and convenient web server for

biologists to obtain the results of their analysis of a given mRNA sequence.

RiboSW is now available on the web at http://RiboSW.mbc.nctu.edu.tw/.

DOI: 10.1261/rna.1623809

PMCID: PMC2704089

PMID: 19460868 [Indexed for MEDLINE]

2244. BMC Bioinformatics. 2009 Jun 16;10 Suppl 6:S4. doi: 10.1186/1471-2105-10-S6-S4.

RetroTector online, a rational tool for analysis of retroviral elements in small

and medium size vertebrate genomic sequences.

Sperber G(1), Lövgren A, Eriksson NE, Benachenhou F, Blomberg J.

Author information:

(1)Physiology unit, Department of Neuroscience, Box 593, Uppsala, Sweden.

goran.sperber@neuro.uu.se

BACKGROUND: The rapid accumulation of genomic information in databases

necessitates rapid and specific algorithms for extracting biologically meaningful

information. More or less complete retroviral sequences, also called proviral or

endogenous retroviral sequences; ERVs, constitutes at least 5% of vertebrate

genomes. After infecting the host, these retroviruses have integrated in germ

line cells, and have then been carried in genomes for at least several 100

million years. A better understanding of structure and function of these

sequences can have profound biological and medical consequences.

METHODS: RetroTector (ReTe) is a platform-independent Java program for

identification and characterization of proviral sequences in vertebrate genomes.

The full ReTe requires a local installation with a MySQL database. Although not

overly complicated, the installation may take some time. A "light" version of

ReTe, (RetroTector online; ROL) which does not require specific installation

procedures is provided, via the World Wide Web.

RESULT: ROL http://www.fysiologi.neuro.uu.se/jbgs/ was implemented under the

Batchelor web interface (A Lövgren et al). It allows both GenBank accession

number, file and FASTA cut-and-paste admission of sequences (5 to 10,000

kilobases). Up to ten submissions can be done simultaneously, allowing batch

analysis of <or= 100 Megabases. Jobs are shown in an IP-number specific list.

Results are text files, and can be viewed with the program, RetroTectorViewer.jar

(at the same site), which has the full graphical capabilities of the basic ReTe

program. A detailed analysis of any retroviral sequences found in the submitted

sequence is graphically presented, exportable in standard formats. With the

current server, a complete analysis of a 1 Megabase sequence is complete in 10

minutes. It is possible to mask nonretroviral repetitive sequences in the

submitted sequence, using host genome specific "brooms", which increase

specificity.

DISCUSSION: Proviral sequences can be hard to recognize, especially if the

integration occurred many million years ago. Precise delineation of LTR, gag,

pro, pol and env can be difficult, requiring manual work. ROL is a way of

simplifying these tasks.

CONCLUSION: ROL provides 1. annotation and presentation of known retroviral

sequences, 2. detection of proviral chains in unknown genomic sequences, with up

to 100 Mbase per submission.

DOI: 10.1186/1471-2105-10-S6-S4

PMCID: PMC2697651

PMID: 19534753 [Indexed for MEDLINE]

2245. BMC Bioinformatics. 2009 Jun 16;10 Suppl 6:S3. doi: 10.1186/1471-2105-10-S6-S3.

Databases of homologous gene families for comparative genomics.

Penel S(1), Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M,

Perrière G.

Author information:

(1)Laboratoire de Biométrie et Biologie Evolutive, CNRS, Université Claude

Bernard - Lyon 1, 43 bd, du 11 Novembre 1918, 69622 Villeurbanne Cedex, France.

penel@biomserv.univ-lyon1.fr

BACKGROUND: Comparative genomics is a central step in many sequence analysis

studies, from gene annotation and the identification of new functional regions in

genomes, to the study of evolutionary processes at the molecular level

(speciation, single gene or whole genome duplications, etc.) and phylogenetics.

In that context, databases providing users high quality homologous families and

sequence alignments as well as phylogenetic trees based on state of the art

algorithms are becoming indispensable.

METHODS: We developed an automated procedure allowing massive all-against-all

similarity searches, gene clustering, multiple alignments computation, and

phylogenetic trees construction and reconciliation. The application of this

procedure to a very large set of sequences is possible through parallel computing

on a large computer cluster.

RESULTS: Three databases were developed using this procedure: HOVERGEN, HOGENOM

and HOMOLENS. These databases share the same architecture but differ in their

content. HOVERGEN contains sequences from vertebrates, HOGENOM is mainly devoted

to completely sequenced microbial organisms, and HOMOLENS is devoted to metazoan

genomes from Ensembl. Access to the databases is provided through Web query

forms, a general retrieval system and a client-server graphical interface. The

later can be used to perform tree-pattern based searches allowing, among other

uses, to retrieve sets of orthologous genes. The three databases, as well as the

software required to build and query them, can be used or downloaded from the

PBIL (Pôle Bioinformatique Lyonnais) site at http://pbil.univ-lyon1.fr/.

DOI: 10.1186/1471-2105-10-S6-S3

PMCID: PMC2697650

PMID: 19534752 [Indexed for MEDLINE]

2246. BMC Bioinformatics. 2009 Jun 16;10 Suppl 6:S18. doi: 10.1186/1471-2105-10-S6-S18.

Annotation and visualization of endogenous retroviral sequences using the

Distributed Annotation System (DAS) and eBioX.

Barrio AM(1), Lagercrantz E, Sperber GO, Blomberg J, Bongcam-Rudloff E.

Author information:

(1)The Linnaeus Centre for Bioinformatics, Uppsala University, Biomedical centre,

P,O, Box 598, SE-75124 Uppsala, Sweden. Alvaro.Martinez.Barrio@lcb.uu.se

BACKGROUND: The Distributed Annotation System (DAS) is a widely used network

protocol for sharing biological information. The distributed aspects of the

protocol enable the use of various reference and annotation servers for

connecting biological sequence data to pertinent annotations in order to depict

an integrated view of the data for the final user.

RESULTS: An annotation server has been devised to provide information about the

endogenous retroviruses detected and annotated by a specialized in silico tool

called RetroTector. We describe the procedure to implement the DAS 1.5 protocol

commands necessary for constructing the DAS annotation server. We use our server

to exemplify those steps. Data distribution is kept separated from visualization

which is carried out by eBioX, an easy to use open source program incorporating

multiple bioinformatics utilities. Some well characterized endogenous

retroviruses are shown in two different DAS clients. A rapid analysis of areas

free from retroviral insertions could be facilitated by our annotations.

CONCLUSION: The DAS protocol has shown to be advantageous in the distribution of

endogenous retrovirus data. The distributed nature of the protocol is also found

to aid in combining annotation and visualization along a genome in order to

enhance the understanding of ERV contribution to its evolution. Reference and

annotation servers are conjointly used by eBioX to provide visualization of ERV

annotations as well as other data sources. Our DAS data source can be found in

the central public DAS service repository, http://www.dasregistry.org, or at

http://loka.bmc.uu.se/das/sources.

DOI: 10.1186/1471-2105-10-S6-S18

PMCID: PMC2697641

PMID: 19534743 [Indexed for MEDLINE]

2247. Bioinformatics. 2009 Jun 15;25(12):i365-73. doi: 10.1093/bioinformatics/btp212.

A partition function algorithm for interacting nucleic acid strands.

Chitsaz H(1), Salari R, Sahinalp SC, Backofen R.

Author information:

(1)Lab for Computational Biology, School of Computing Science, Simon Fraser

University, Burnaby, BC, Canada.

Recent interests, such as RNA interference and antisense RNA regulation, strongly

motivate the problem of predicting whether two nucleic acid strands

interact.MOTIVATION: Regulatory non-coding RNAs (ncRNAs) such as microRNAs play

an important role in gene regulation. Studies on both prokaryotic and eukaryotic

cells show that such ncRNAs usually bind to their target mRNA to regulate the

translation of corresponding genes. The specificity of these interactions depends

on the stability of intermolecular and intramolecular base pairing. While methods

like deep sequencing allow to discover an ever increasing set of ncRNAs, there

are no high-throughput methods available to detect their associated targets.

Hence, there is an increasing need for precise computational target prediction.

In order to predict base-pairing probability of any two bases in interacting

nucleic acids, it is necessary to compute the interaction partition function over

the whole ensemble. The partition function is a scalar value from which various

thermodynamic quantities can be derived. For example, the equilibrium

concentration of each complex nucleic acid species and also the melting

temperature of interacting nucleic acids can be calculated based on the partition

function of the complex.

RESULTS: We present a model for analyzing the thermodynamics of two interacting

nucleic acid strands considering the most general type of interactions studied in

the literature. We also present a corresponding dynamic programming algorithm

that computes the partition function over (almost) all physically possible joint

secondary structures formed by two interacting nucleic acids in O(n(6)) time. We

verify the predictive power of our algorithm by computing (i) the melting

temperature for interacting RNA pairs studied in the literature and (ii) the

equilibrium concentration for several variants of the OxyS-fhlA complex. In both

experiments, our algorithm shows high accuracy and outperforms competitors.

AVAILABILITY: Software and web server is available at

http://compbio.cs.sfu.ca/taverna/pirna/.

SUPPLEMENTARY INFORMATION: Supplementary data are avaliable at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp212

PMCID: PMC2687966

PMID: 19478011 [Indexed for MEDLINE]

2248. Bioinformatics. 2009 Jun 15;25(12):i179-86. doi: 10.1093/bioinformatics/btp223.

E-zyme: predicting potential EC numbers from the chemical transformation pattern

of substrate-product pairs.

Yamanishi Y(1), Hattori M, Kotera M, Goto S, Kanehisa M.

Author information:

(1)Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji,

Kyoto, Japan.

MOTIVATION: The IUBMB's Enzyme Nomenclature system, commonly known as the Enzyme

Commission (EC) numbers, plays key roles in classifying enzymatic reactions and

in linking the enzyme genes or proteins to reactions in metabolic pathways. There

are numerous reactions known to be present in various pathways but without any

official EC numbers, most of which have no hope to be given ones because of the

lack of the published articles on enzyme assays.

RESULTS: In this article we propose a new method to predict the potential EC

numbers to given reactant pairs (substrates and products) or uncharacterized

reactions, and a web-server named E-zyme as an application. This technology is

based on our original biochemical transformation pattern which we call an 'RDM

pattern', and consists of three steps: (i) graph alignment of a query reactant

pair (substrates and products) for computing the query RDM pattern, (ii)

multi-layered partial template matching by comparing the query RDM pattern with

template patterns related with known EC numbers and (iii) weighted major voting

scheme for selecting appropriate EC numbers. As the result, cross-validation

experiments show that the proposed method achieves both high coverage and high

prediction accuracy at a practical level, and consistently outperforms the

previous method.

AVAILABILITY: The E-zyme system is available at

http://www.genome.jp/tools/e-zyme/.

DOI: 10.1093/bioinformatics/btp223

PMCID: PMC2687977

PMID: 19477985 [Indexed for MEDLINE]

2249. Bioinformatics. 2009 Jun 15;25(12):1552-3. doi: 10.1093/bioinformatics/btp248.

Epub 2009 Apr 23.

CROC: finding chromosomal clusters in eukaryotic genomes.

Pignatelli M(1), Serras F, Moya A, Guigó R, Corominas M.

Author information:

(1)Instituto Cavanilles of Biodiversity and Evolutionary Biology, University of

Valencia, Apdo, Valencia, Spain. miguel.pignatelli@uv.es

SUMMARY: There is increasing evidence showing that co-expression of genes that

cluster along the genome is a common characteristic of eukaryotic transcriptomes.

Several algorithms have been used to date in the identification of these kinds of

gene organization. Here, we present a web tool called CROC that aims to help in

the identification and analysis of genomic gene clusters. This method has been

successfully used before in the identification of chromosomal clusters in

different eukaryotic species.

AVAILABILITY: The web server is freely available to non-commercial users at the

following address: http://metagenomics.uv.es/CROC/.

DOI: 10.1093/bioinformatics/btp248

PMID: 19389737 [Indexed for MEDLINE]

2250. Bioinformatics. 2009 Jun 15;25(12):1506-12. doi: 10.1093/bioinformatics/btp238.

Epub 2009 Apr 8.

ModLink+: improving fold recognition by using protein-protein interactions.

Fornes O(1), Aragues R, Espadaler J, Marti-Renom MA, Sali A, Oliva B.

Author information:

(1)Structural Bioinformatics Lab (GRIB-IMIM), Universitat Pompeu Fabra, Parc de

Recerca Biomèdica de Barcelona (PRBB), Barcelona, Catalonia, Spain.

boliva@imim.es

MOTIVATION: Several strategies have been developed to predict the fold of a

target protein sequence, most of which are based on aligning the target sequence

to other sequences of known structure. Previously, we demonstrated that the

consideration of protein-protein interactions significantly increases the

accuracy of fold assignment compared with PSI-BLAST sequence comparisons. A

drawback of our method was the low number of proteins to which a fold could be

assigned. Here, we present an improved version of the method that addresses this

limitation. We also compare our method to other state-of-the-art fold assignment

methodologies.

RESULTS: Our approach (ModLink+) has been tested on 3716 proteins with domain

folds classified in the Structural Classification Of Proteins (SCOP) as well as

known interacting partners in the Database of Interacting Proteins (DIP). For

this test set, the ratio of success [positive predictive value (PPV)] on fold

assignment increases from 75% for PSI-BLAST, 83% for HHSearch and 81% for PRC to

>90% for ModLink+at the e-value cutoff of 10(-3). Under this e-value, ModLink+can

assign a fold to 30-45% of the proteins in the test set, while our previous

method could cover <25%. When applied to 6384 proteins with unknown fold in the

yeast proteome, ModLink+combined with PSI-BLAST assigns a fold for domains in

3738 proteins, while PSI-BLAST alone covers only 2122 proteins, HHSearch 2969 and

PRC 2826 proteins, using a threshold e-value that would represent a PPV >82% for

each method in the test set.

AVAILABILITY: The ModLink+server is freely accessible in the World Wide Web at

http://sbi.imim.es/modlink/.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp238

PMCID: PMC2687990

PMID: 19357100 [Indexed for MEDLINE]

2251. Bioinformatics. 2009 Jun 15;25(12):1550-1. doi: 10.1093/bioinformatics/btp239.

Epub 2009 Apr 8.

WebGBrowse--a web server for GBrowse.

Podicheti R(1), Gollapudi R, Dong Q.

Author information:

(1)Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN

47405, USA.

SUMMARY: The Generic Genome Browser (GBrowse) is one of the most widely used

tools for visualizing genomic features along a reference sequence. However, the

installation and configuration of GBrowse is not trivial for biologists. We have

developed a web server, WebGBrowse that allows users to upload genome annotation

in the GFF3 format, configure the display of each genomic feature by simply using

a web browser and visualize the configured genomic features with the integrated

GBrowse software.

AVAILABILITY: WebGBrowse is accessible via http://webgbrowse.cgb.indiana.edu/ and

the system is also freely available for local installations.

DOI: 10.1093/bioinformatics/btp239

PMID: 19357095 [Indexed for MEDLINE]

2252. PLoS One. 2009 Jun 15;4(6):e5917. doi: 10.1371/journal.pone.0005917.

Prediction of type III secretion signals in genomes of gram-negative bacteria.

Löwer M(1), Schneider G.

Author information:

(1)Johann Wolfgang Goethe-University, Chair for Chem- and Bioinformatics,

Frankfurt, Germany.

Erratum in

PLoS One. 2009;4(7). doi:

10.1371/annotation/78c8fc32-b1e2-4c87-9c92-d318af980b9b.

BACKGROUND: Pathogenic bacteria infecting both animals as well as plants use

various mechanisms to transport virulence factors across their cell membranes and

channel these proteins into the infected host cell. The type III secretion system

represents such a mechanism. Proteins transported via this pathway ("effector

proteins") have to be distinguished from all other proteins that are not exported

from the bacterial cell. Although a special targeting signal at the N-terminal

end of effector proteins has been proposed in literature its exact

characteristics remain unknown.

METHODOLOGY/PRINCIPAL FINDINGS: In this study, we demonstrate that the signals

encoded in the sequences of type III secretion system effectors can be

consistently recognized and predicted by machine learning techniques. Known

protein effectors were compiled from the literature and sequence databases, and

served as training data for artificial neural networks and support vector machine

classifiers. Common sequence features were most pronounced in the first 30 amino

acids of the effector sequences. Classification accuracy yielded a

cross-validated Matthews correlation of 0.63 and allowed for genome-wide

prediction of potential type III secretion system effectors in 705

proteobacterial genomes (12% predicted candidates protein), their chromosomes

(11%) and plasmids (13%), as well as 213 Firmicute genomes (7%).

CONCLUSIONS/SIGNIFICANCE: We present a signal prediction method together with

comprehensive survey of potential type III secretion system effectors extracted

from 918 published bacterial genomes. Our study demonstrates that the analyzed

signal features are common across a wide range of species, and provides a

substantial basis for the identification of exported pathogenic proteins as

targets for future therapeutic intervention. The prediction software is publicly

accessible from our web server (www.modlab.org).

DOI: 10.1371/journal.pone.0005917

PMCID: PMC2690842

PMID: 19526054 [Indexed for MEDLINE]

2253. Bioinform Biol Insights. 2009 Jun 3;3:51-69.

Predicting consensus structures for RNA alignments via pseudo-energy

minimization.

Spirollari J(1), Wang JT, Zhang K, Bellofatto V, Park Y, Shapiro BA.

Author information:

(1)Bioinformatics Program, Department of Computer Science, New Jersey Institute

of Technology, University Heights, Newark, NJ 07102, U.S.A.

Thermodynamic processes with free energy parameters are often used in algorithms

that solve the free energy minimization problem to predict secondary structures

of single RNA sequences. While results from these algorithms are promising, an

observation is that single sequence-based methods have moderate accuracy and more

information is needed to improve on RNA secondary structure prediction, such as

covariance scores obtained from multiple sequence alignments. We present in this

paper a new approach to predicting the consensus secondary structure of a set of

aligned RNA sequences via pseudo-energy minimization. Our tool, called RSpredict,

takes into account sequence covariation and employs effective heuristics for

accuracy improvement. RSpredict accepts, as input data, a multiple sequence

alignment in FASTA or ClustalW format and outputs the consensus secondary

structure of the input sequences in both the Vienna style Dot Bracket format and

the Connectivity Table format. Our method was compared with some widely used

tools including KNetFold, Pfold and RNAalifold. A comprehensive test on different

datasets including Rfam sequence alignments and a multiple sequence alignment

obtained from our study on the Drosophila X chromosome reveals that RSpredict is

competitive with the existing tools on the tested datasets. RSpredict is freely

available online as a web server and also as a jar file for download at

http://datalab.njit.edu/biology/RSpredict.

PMCID: PMC2808183

PMID: 20140072

2254. Bioinformatics. 2009 Jun 1;25(11):1461-2. doi: 10.1093/bioinformatics/btp176.

Epub 2009 Apr 17.

FDR made easy in differential feature discovery and correlation analyses.

Ling XB(1), Cohen H, Jin J, Lau I, Schilling J.

Author information:

(1)Department of Pediatrics, Stanford University School of Medicine, Stanford

University, Stanford, CA 94305, USA. xuefeng\_ling@yahoo.com

SUMMARY: Rapid progress in technology, particularly in high-throughput biology,

allows the analysis of thousands of genes or proteins simultaneously, where the

multiple comparison problems occurs. Global false discovery rate (gFDR) analysis

statistically controls this error, computing the ratio of the number of false

positives over the total number of rejections. Local FDR (lFDR) method can

associate the corrected significance measure with each hypothesis testing for its

feature-by-feature interpretation. Given the large feature number and sample size

in any genomics or proteomics analysis, FDR computation, albeit critical, is both

beyond the regular biologists' specialty and computationally expensive, easily

exceeding the capacity of desktop computers. To overcome this digital divide, a

web portal has been developed that provides bench-side biologists easy access to

the server-side computing capabilities to analyze for FDR, differential expressed

genes or proteins, and for the correlation between molecular data and clinical

measurements.

AVAILABILITY:

(http://translationalmedicine.stanford.edu/Mass-Conductor/FDR.html).

DOI: 10.1093/bioinformatics/btp176

PMID: 19376824 [Indexed for MEDLINE]

2255. Bioinformatics. 2009 Jun 1;25(11):1433-4. doi: 10.1093/bioinformatics/btp251.

Epub 2009 Apr 9.

MICAlign: a sequence-to-structure alignment tool integrating multiple sources of

information in conditional random fields.

Xia X(1), Zhang S, Su Y, Sun Z.

Author information:

(1)Department of Biological Sciences and Biotechnology, MOE Key Laboratory of

Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology,

Tsinghua University, Beijing, China.

SUMMARY: Sequence-to-structure alignment in template-based protein structure

modeling for remote homologs remains a difficult problem even following the

correct recognition of folds. Here we present MICAlign, a sequence-to-structure

alignment tool that incorporates multiple sources of information from local

structural contexts of template, sequence profiles, predicted secondary

structures, solvent accessibilities, potential-like terms (including

residue-residue contacts and solvent exposures) and pre-aligned structures and

sequences. These features, together with a position-specific gap scheme, were

integrated into conditional random fields through which the optimal parameters

were automatically learned. MICAlign showed improved alignment accuracy over

several other state-of-the-art alignment tools based on comparisons by using

independent datasets.

AVAILABILITY: Freely available at

(http://www.bioinfo.tsinghua.edu.cn/~xiaxf/micalign) for both web server and

source code.

DOI: 10.1093/bioinformatics/btp251

PMID: 19359356 [Indexed for MEDLINE]

2256. Bioinformatics. 2009 Jun 1;25(11):1438-9. doi: 10.1093/bioinformatics/btp165.

Epub 2009 Mar 23.

BioconductorBuntu: a Linux distribution that implements a web-based DNA

microarray analysis server.

Geeleher P(1), Morris D, Hinde JP, Golden A.

Author information:

(1)School of Mathematics, Statistics and Applied Mathematics, National University

of Ireland, Galway, Ireland. paul.geeleher@nuigalway.ie

SUMMARY: BioconductorBuntu is a custom distribution of Ubuntu Linux that

automatically installs a server-side microarray processing environment, providing

a user-friendly web-based GUI to many of the tools developed by the Bioconductor

Project, accessible locally or across a network. System installation is via

booting off a CD image or by using a Debian package provided to upgrade an

existing Ubuntu installation. In its current version, several microarray analysis

pipelines are supported including oligonucleotide, dual-or single-dye

experiments, including post-processing with Gene Set Enrichment Analysis.

BioconductorBuntu is designed to be extensible, by server-side integration of

further relevant Bioconductor modules as required, facilitated by its

straightforward underlying Python-based infrastructure. BioconductorBuntu offers

an ideal environment for the development of processing procedures to facilitate

the analysis of next-generation sequencing datasets.

AVAILABILITY: BioconductorBuntu is available for download under a creative

commons license along with additional documentation and a tutorial from

(http://bioinf.nuigalway.ie).

DOI: 10.1093/bioinformatics/btp165

PMID: 19307241 [Indexed for MEDLINE]

2257. Bioinformatics. 2009 Jun 1;25(11):1426-7. doi: 10.1093/bioinformatics/btp160.

Epub 2009 Mar 23.

Evolutionary Trace Annotation Server: automated enzyme function prediction in

protein structures using 3D templates.

Ward RM(1), Venner E, Daines B, Murray S, Erdin S, Kristensen DM, Lichtarge O.

Author information:

(1)Department of Molecular and Human Genetics, Program in Structural and

Computational Biology and Molecular, Baylor College of Medicine, One Baylor

Plaza, Houston, TX 77030, USA.

SUMMARY: The Evolutionary Trace Annotation (ETA) Server predicts enzymatic

activity. ETA starts with a structure of unknown function, such as those from

structural genomics, and with no prior knowledge of its mechanism uses the

phylogenetic Evolutionary Trace (ET) method to extract key functional residues

and propose a function-associated 3D motif, called a 3D template. ETA then

searches previously annotated structures for geometric template matches that

suggest molecular and thus functional mimicry. In order to maximize the

predictive value of these matches, ETA next applies distinctive specificity

filters -- evolutionary similarity, function plurality and match reciprocity. In

large scale controls on enzymes, prediction coverage is 43% but the positive

predictive value rises to 92%, thus minimizing false annotations. Users may

modify any search parameter, including the template. ETA thus expands the ET

suite for protein structure annotation, and can contribute to the annotation

efforts of metaservers.

AVAILABILITY: The ETA Server is a web application available at

(http://mammoth.bcm.tmc.edu/eta/).

DOI: 10.1093/bioinformatics/btp160

PMCID: PMC2682511

PMID: 19307237 [Indexed for MEDLINE]

2258. J Digit Imaging. 2009 Jun;22(3):250-8. Epub 2007 Nov 13.

The application of JPEG2000 in virtual microscopy.

Tuominen VJ(1), Isola J.

Author information:

(1)Institute of Medical Technology, Tampere University Hospital, University of

Tampere, Tampere, Finland. vilppu.tuominen@iki.fi

Virtual microscopy (i.e., the viewing of entire microscope specimens on a

computer display) is becoming widely applied in microscopy teaching and clinical

laboratory medicine. Despite rapidly increasing use, virtual microscopy currently

lacks of a universally accepted image format. A promising candidate is JPEG2000,

which has potential advantages for handling gigabyte-sized virtual slides. To

date, no JPEG2000-based software has been specifically suited for virtual

microscopy. To study the utility of JPEG2000 in virtual microscopy, we first

optimized JPEG2000 code-stream parameters for virtual slide viewing (i.e., fast

navigation, zooming, and use of an overview window). Compression using ratios

25:1-30:1 with the irreversible wavelet filter were found to provide the best

compromise between file size and image quality. Optimal code-stream parameters

also consisted of 10 wavelet decomposition levels, progression order

Resolution-Position-Component-Layer (RPCL), a precinct size of 128 x 128, and

code-block size of 64 x 64. Tiling and the use of multiple quality layers were

deemed unnecessary. A compression application (JVScomp) was developed for

creating optimally parameterized JPEG2000 virtual slides. A viewing application

(JVSview) was developed specifically for virtual microscopy, offering all of the

basic viewing functions. JVSview also supports viewing of focus stacks, embedding

of textual descriptions, and defining regions of interest as metadata. Combined

with our server application (JVSserv), virtual slides can be viewed over networks

by employing the JPEG2000 Interactive Protocol (JPIP). The software can be tested

using virtual slide examples located on our public JPIP server (

http://jvsmicroscope.uta.fi/ ). The software package is freely downloadable and

usable for noncommercial purposes.

DOI: 10.1007/s10278-007-9090-z

PMCID: PMC3043697

PMID: 17999112 [Indexed for MEDLINE]

2259. Nucleic Acids Res. 2009 Jun;37(11):3522-30. doi: 10.1093/nar/gkp212. Epub 2009

Apr 7.

PROCAIN: protein profile comparison with assisting information.

Wang Y(1), Sadreyev RI, Grishin NV.

Author information:

(1)Biomedical Engineering Program, University of Texas Southwestern Medical

Center, Dallas, TX 75390-9050, USA.

Detection of remote sequence homology is essential for the accurate inference of

protein structure, function and evolution. The most sensitive detection methods

involve the comparison of evolutionary patterns reflected in multiple sequence

alignments (MSAs) of protein families. We present PROCAIN, a new method for MSA

comparison based on the combination of 'vertical' MSA context (substitution

constraints at individual sequence positions) and 'horizontal' context (patterns

of residue content at multiple positions). Based on a simple and tractable

profile methodology and primitive measures for the similarity of horizontal MSA

patterns, the method achieves the quality of homology detection comparable to a

more complex advanced method employing hidden Markov models (HMMs) and secondary

structure (SS) prediction. Adding SS information further improves PROCAIN

performance beyond the capabilities of current state-of-the-art tools. The

potential value of the method for structure/function predictions is illustrated

by the detection of subtle homology between evolutionary distant yet structurally

similar protein domains. ProCAIn, relevant databases and tools can be downloaded

from: http://prodata.swmed.edu/procain/download. The web server can be accessed

at http://prodata.swmed.edu/procain/procain.php.

DOI: 10.1093/nar/gkp212

PMCID: PMC2699500

PMID: 19357092 [Indexed for MEDLINE]

2260. BMC Bioinformatics. 2009 May 26;10:159. doi: 10.1186/1471-2105-10-159.

Transmembrane protein topology prediction using support vector machines.

Nugent T(1), Jones DT.

Author information:

(1)Bioinformatics Group, Department of Computer Science, University College

London, Gower Street, London, WC1E 6BT, UK. t.nugent@cs.ucl.ac.uk

BACKGROUND: Alpha-helical transmembrane (TM) proteins are involved in a wide

range of important biological processes such as cell signaling, transport of

membrane-impermeable molecules, cell-cell communication, cell recognition and

cell adhesion. Many are also prime drug targets, and it has been estimated that

more than half of all drugs currently on the market target membrane proteins.

However, due to the experimental difficulties involved in obtaining high quality

crystals, this class of protein is severely under-represented in structural

databases. In the absence of structural data, sequence-based prediction methods

allow TM protein topology to be investigated.

RESULTS: We present a support vector machine-based (SVM) TM protein topology

predictor that integrates both signal peptide and re-entrant helix prediction,

benchmarked with full cross-validation on a novel data set of 131 sequences with

known crystal structures. The method achieves topology prediction accuracy of

89%, while signal peptides and re-entrant helices are predicted with 93% and 44%

accuracy respectively. An additional SVM trained to discriminate between globular

and TM proteins detected zero false positives, with a low false negative rate of

0.4%. We present the results of applying these tools to a number of complete

genomes. Source code, data sets and a web server are freely available from

http://bioinf.cs.ucl.ac.uk/psipred/.

CONCLUSION: The high accuracy of TM topology prediction which includes detection

of both signal peptides and re-entrant helices, combined with the ability to

effectively discriminate between TM and globular proteins, make this method

ideally suited to whole genome annotation of alpha-helical transmembrane

proteins.

DOI: 10.1186/1471-2105-10-159

PMCID: PMC2700806

PMID: 19470175 [Indexed for MEDLINE]

2261. BMC Evol Biol. 2009 May 18;9:108. doi: 10.1186/1471-2148-9-108.

PhyloExplorer: a web server to validate, explore and query phylogenetic trees.

Ranwez V(1), Clairon N, Delsuc F, Pourali S, Auberval N, Diser S, Berry V.

Author information:

(1)Institut des Sciences de l'Evolution (ISEM, UMR 5554 CNRS), Université

Montpellier II, Place E, Bataillon - 34095 Montpellier Cedex 05, France.

vincent.ranwez@univ-montp2.fr

BACKGROUND: Many important problems in evolutionary biology require molecular

phylogenies to be reconstructed. Phylogenetic trees must then be manipulated for

subsequent inclusion in publications or analyses such as supertree inference and

tree comparisons. However, no tool is currently available to facilitate the

management of tree collections providing, for instance: standardisation of taxon

names among trees with respect to a reference taxonomy; selection of relevant

subsets of trees or sub-trees according to a taxonomic query; or simply

computation of descriptive statistics on the collection. Moreover, although

several databases of phylogenetic trees exist, there is currently no easy way to

find trees that are both relevant and complementary to a given collection of

trees.

RESULTS: We propose a tool to facilitate assessment and management of

phylogenetic tree collections. Given an input collection of rooted trees,

PhyloExplorer provides facilities for obtaining statistics describing the

collection, correcting invalid taxon names, extracting taxonomically relevant

parts of the collection using a dedicated query language, and identifying related

trees in the TreeBASE database.

CONCLUSION: PhyloExplorer is a simple and interactive website implemented through

underlying Python libraries and MySQL databases. It is available at:

http://www.ncbi.orthomam.univ-montp2.fr/phyloexplorer/ and the source code can be

downloaded from: http://code.google.com/p/taxomanie/.

DOI: 10.1186/1471-2148-9-108

PMCID: PMC2695458

PMID: 19450253 [Indexed for MEDLINE]

2262. Bioinformatics. 2009 May 15;25(10):1259-63. doi: 10.1093/bioinformatics/btp148.

Epub 2009 Mar 25.

Structure similarity measure with penalty for close non-equivalent residues.

Sadreyev RI(1), Shi S, Baker D, Grishin NV.

Author information:

(1)Howard Hughes Medical Institute, University of Texas Southwestern Medical

Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA.

MOTIVATION: Recent improvement in homology-based structure modeling emphasizes

the importance of sensitive evaluation measures that help identify and correct

modest distortions in models compared with the target structures. Global Distance

Test Total Score (GDT\_TS), otherwise a very powerful and effective measure for

model evaluation, is still insensitive to and can even reward such distortions,

as observed for remote homology modeling in the latest CASP8 (Comparative

Assessment of Structure Prediction).

RESULTS: We develop a new measure that balances GDT\_TS reward for the closeness

of equivalent model and target residues ('attraction' term) with the penalty for

the closeness of non-equivalent residues ('repulsion' term). Compared with

GDT\_TS, the resulting score, TR (total score with repulsion), is much more

sensitive to structure compression both in real remote homologs and in CASP

models. TR is correlated yet different from other measures of structure

similarity. The largest difference from GDT\_TS is observed in models of mid-range

quality based on remote homology modeling.

AVAILABILITY: The script for TR calculation is included in Supplementary

Material. TR scores for all server models in CASP8 are available at

http://prodata.swmed.edu/CASP8.

DOI: 10.1093/bioinformatics/btp148

PMCID: PMC2677741

PMID: 19321733 [Indexed for MEDLINE]

2263. Bioinformatics. 2009 May 15;25(10):1331-2. doi: 10.1093/bioinformatics/btp141.

Epub 2009 Mar 11.

CodonExplorer: an online tool for analyzing codon usage and sequence composition,

scaling from genes to genomes.

Hamady M(1), Wilson SA, Zaneveld J, Sueoka N, Knight R.

Author information:

(1)Department of Computer Science, University of Colorado, Boulder, CO 80309,

USA.

DNA composition in general, and codon usage in particular, is crucial for

understanding gene function and evolution. CodonExplorer, available online at

http://bmf.colorado.edu/codonexplorer/, is an online tool and interactive

database that contains millions of genes, allowing rapid exploration of the

factors governing gene and genome compositional evolution and exploiting GC

content and codon usage frequency to identify genes with composition suggesting

high levels of expression or horizontal transfer.

DOI: 10.1093/bioinformatics/btp141

PMCID: PMC2677738

PMID: 19279067 [Indexed for MEDLINE]

2264. BMC Struct Biol. 2009 May 13;9:30. doi: 10.1186/1472-6807-9-30.

Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins

using neural networks.

Andrabi M(1), Mizuguchi K, Sarai A, Ahmad S.

Author information:

(1)National Institute of Biomedical Innovation, Ibaraki-shi, Osaka, Japan.

munazah@nibio.go.jp

BACKGROUND: DNA recognition by proteins is one of the most important processes in

living systems. Therefore, understanding the recognition process in general, and

identifying mutual recognition sites in proteins and DNA in particular, carries

great significance. The sequence and structural dependence of DNA-binding sites

in proteins has led to the development of successful machine learning methods for

their prediction. However, all existing machine learning methods predict

DNA-binding sites, irrespective of their target sequence and hence, none of them

is helpful in identifying specific protein-DNA contacts. In this work, we

formulate the problem of predicting specific DNA-binding sites in terms of

contacts between the residue environments of proteins and the identity of a

mononucleotide or a dinucleotide step in DNA. The aim of this work is to take a

protein sequence or structural features as inputs and predict for each amino acid

residue if it binds to DNA at locations identified by one of the four possible

mononucleotides or one of the 10 unique dinucleotide steps. Contact predictions

are made at various levels of resolution viz. in terms of side chain, backbone

and major or minor groove atoms of DNA.

RESULTS: Significant differences in residue preferences for specific contacts are

observed, which combined with other features, lead to promising levels of

prediction. In general, PSSM-based predictions, supported by secondary structure

and solvent accessibility, achieve a good predictability of approximately 70-80%,

measured by the area under the curve (AUC) of ROC graphs. The major and minor

groove contact predictions stood out in terms of their poor predictability from

sequences or PSSM, which was very strongly (>20 percentage points) compensated by

the addition of secondary structure and solvent accessibility information,

revealing a predominant role of local protein structure in the major/minor groove

DNA-recognition. Following a detailed analysis of results, a web server to

predict mononucleotide and dinucleotide-step contacts using PSSM was developed

and made available at http://sdcpred.netasa.org/ or

http://tardis.nibio.go.jp/netasa/sdcpred/.

CONCLUSION: Most residue-nucleotide contacts can be predicted with high accuracy

using only sequence and evolutionary information. Major and minor groove

contacts, however, depend profoundly on the local structure. Overall, this study

takes us a step closer to the ultimate goal of predicting mutual recognition

sites in protein and DNA sequences.

DOI: 10.1186/1472-6807-9-30

PMCID: PMC2693520

PMID: 19439068 [Indexed for MEDLINE]

2265. Bioinformatics. 2009 May 1;25(9):1185-6. doi: 10.1093/bioinformatics/btp121. Epub

2009 Mar 4.

A web server for inferring the human N-acetyltransferase-2 (NAT2) enzymatic

phenotype from NAT2 genotype.

Kuznetsov IB(1), McDuffie M, Moslehi R.

Author information:

(1)Gen\*NY\*Sis Center for Excellence in Cancer Genomics, Department of

Epidemiology and Biostatistics, University at Albany, One Discovery Drive,

Rensselaer, NY 12144, USA. ikuznetsov@albany.edu

N-acetyltransferase-2 (NAT2) is an important enzyme that catalyzes the

acetylation of aromatic and heterocyclic amine carcinogens. Individuals in human

populations are divided into three NAT2 acetylator phenotypes: slow, rapid and

intermediate. NAT2PRED is a web server that implements a supervised pattern

recognition method to infer NAT2 phenotype from SNPs found in NAT2 gene positions

282, 341, 481, 590, 803 and 857. The web server can be used for a fast

determination of NAT2 phenotypes in genetic screens.AVAILABILITY: Freely

available at http://nat2pred.rit.albany.edu.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp121

PMCID: PMC2672629

PMID: 19261719 [Indexed for MEDLINE]

2266. Bioinformatics. 2009 May 1;25(9):1189-91. doi: 10.1093/bioinformatics/btp033.

Epub 2009 Jan 16.

Jalview Version 2--a multiple sequence alignment editor and analysis workbench.

Waterhouse AM(1), Procter JB, Martin DM, Clamp M, Barton GJ.

Author information:

(1)School of Life Sciences Research, College of Life Sciences, University of

Dundee, Dow Street, Dundee DD1 5EH, UK.

Jalview Version 2 is a system for interactive WYSIWYG editing, analysis and

annotation of multiple sequence alignments. Core features include keyboard and

mouse-based editing, multiple views and alignment overviews, and linked structure

display with Jmol. Jalview 2 is available in two forms: a lightweight Java applet

for use in web applications, and a powerful desktop application that employs web

services for sequence alignment, secondary structure prediction and the retrieval

of alignments, sequences, annotation and structures from public databases and any

DAS 1.53 compliant sequence or annotation server.AVAILABILITY: The Jalview 2

Desktop application and JalviewLite applet are made freely available under the

GPL, and can be downloaded from www.jalview.org.

DOI: 10.1093/bioinformatics/btp033

PMCID: PMC2672624

PMID: 19151095 [Indexed for MEDLINE]

2267. Diagn Cytopathol. 2009 May;37(5):340-6. doi: 10.1002/dc.20996.

Use of streamed internet video for cytology training and education:

www.PathLab.org.

Poller D(1), Ljung BM, Gonda P.

Author information:

(1)Department of Pathology, Queen Alexandra Hospital, Portsmouth, United Kingdom.

david.poller@porthosp.nhs.uk

An Internet-based method is described for submission of video clips to a website

editor to be reviewed, edited, and then uploaded onto a video server, with a

hypertext link to a website. The information on the webpages is searchable via

the website sitemap on Internet search engines. A survey of video users who

accessed a single 59-minute FNA cytology training cytology video via the website

showed a mean score for usefulness for specialists/consultants of 3.75, range

1-5, n = 16, usefulness for trainees mean score was 4.4, range 3-5, n = 12, with

a mean score for visual and sound quality of 3.9, range 2-5, n = 16. Fifteen out

of 17 respondents thought that posting video training material on the Internet

was a good idea, and 9 of 17 respondents would also consider submitting training

videos to a similar website. This brief exercise has shown that there is value in

posting educational or training video content on the Internet and that the use of

streamed video accessed via the Internet will be of increasing importance.

(c) 2009 Wiley-Liss, Inc.

DOI: 10.1002/dc.20996

PMID: 19191291 [Indexed for MEDLINE]

2268. J Genet Genomics. 2009 May;36(5):289-96. doi: 10.1016/S1673-8527(08)60117-4.

BiodMHC: an online server for the prediction of MHC class II-peptide binding

affinity.

Wang L(1), Pan D, Hu X, Xiao J, Gao Y, Zhang H, Zhang Y, Liu J, Zhu S.

Author information:

(1)School of Computer, Wuhan University, Wuhan 430079, China.

Effective identification of major histocompatibility complex (MHC) molecules

restricted peptides is a critical step in discovering immune epitopes. Although

many online servers have been built to predict class II MHC-peptide binding

affinity, they have been trained on different datasets, and thus fail in

providing a unified comparison of various methods. In this paper, we present our

implementation of seven popular predictive methods, namely SMM-align, ARB,

SVR-pairwise, Gibbs sampler, ProPred, LP-top2, and MHCPred, on a single web

server named BiodMHC (http://biod.whu.edu.cn/BiodMHC/index.html, the software is

available upon request). Using a standard measure of AUC (Area Under the receiver

operating characteristic Curves), we compare these methods by means of not only

cross validation but also prediction on independent test datasets. We find that

SMM-align, ProPred, SVR-pairwise, ARB, and Gibbs sampler are the five

best-performing methods. For the binding affinity prediction of class II

MHC-peptide, BiodMHC provides a convenient online platform for researchers to

obtain binding information simultaneously using various methods.

DOI: 10.1016/S1673-8527(08)60117-4

PMID: 19447377 [Indexed for MEDLINE]

2269. BMC Genomics. 2009 Apr 21;10:174. doi: 10.1186/1471-2164-10-174.

ESTPiper--a web-based analysis pipeline for expressed sequence tags.

Tang Z(1), Choi JH, Hemmerich C, Sarangi A, Colbourne JK, Dong Q.

Author information:

(1)The Center for Genomics and Bioinformatics, Indiana University, Bloomington,

Indiana, USA. tangz@cgb.indiana.edu

BACKGROUND: EST sequencing projects are increasing in scale and scope as the

genome sequencing technologies migrate from core sequencing centers to individual

research laboratories. Effectively, generating EST data is no longer a bottleneck

for investigators. However, processing large amounts of EST data remains a

non-trivial challenge for many. Web-based EST analysis tools are proving to be

the most convenient option for biologists when performing their analysis, so

these tools must continuously improve on their utility to keep in step with the

growing needs of research communities. We have developed a web-based EST analysis

pipeline called ESTPiper, which streamlines typical large-scale EST analysis

components.

RESULTS: The intuitive web interface guides users through each step of base

calling, data cleaning, assembly, genome alignment, annotation, analysis of gene

ontology (GO), and microarray oligonucleotide probe design. Each step is

modularized. Therefore, a user can execute them separately or together in batch

mode. In addition, the user has control over the parameters used by the

underlying programs. Extensive documentation of ESTPiper's functionality is

embedded throughout the web site to facilitate understanding of the required

input and interpretation of the computational results. The user can also download

intermediate results and port files to separate programs for further analysis. In

addition, our server provides a time-stamped description of the run history for

reproducibility. The pipeline can also be installed locally, allowing researchers

to modify ESTPiper to suit their own needs.

CONCLUSION: ESTPiper streamlines the typical process of EST analysis. The

pipeline was initially designed in part to support the Daphnia pulex cDNA

sequencing project. A web server hosting ESTPiper is provided at

http://estpiper.cgb.indiana.edu/ to now support projects of all size. The

software is also freely available from the authors for local installations.

DOI: 10.1186/1471-2164-10-174

PMCID: PMC2676306

PMID: 19383159 [Indexed for MEDLINE]

2270. BMC Bioinformatics. 2009 Apr 9;10:105. doi: 10.1186/1471-2105-10-105.

Prediction of guide strand of microRNAs from its sequence and secondary

structure.

Ahmed F(1), Ansari HR, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India.

firoz@imtech.res.in

BACKGROUND: MicroRNAs (miRNAs) are produced by the sequential processing of a

long hairpin RNA transcript by Drosha and Dicer, an RNase III enzymes, and form

transitory small RNA duplexes. One strand of the duplex, which incorporates into

RNA-induced silencing complex (RISC) and silences the gene expression is called

guide strand, or miRNA; while the other strand of duplex is degraded and called

the passenger strand, or miRNA\*. Predicting the guide strand of miRNA is

important for better understanding the RNA interference pathways.

RESULTS: This paper describes support vector machine (SVM) models developed for

predicting the guide strands of miRNAs. All models were trained and tested on a

dataset consisting of 329 miRNA and 329 miRNA\* pairs using five fold cross

validation technique. Firstly, models were developed using mono-, di-, and

tri-nucleotide composition of miRNA strands and achieved the highest accuracies

of 0.588, 0.638 and 0.596 respectively. Secondly, models were developed using

split nucleotide composition and achieved maximum accuracies of 0.553, 0.641 and

0.602 for mono-, di-, and tri-nucleotide respectively. Thirdly, models were

developed using binary pattern and achieved the highest accuracy of 0.708.

Furthermore, when integrating the secondary structure features with binary

pattern, an accuracy of 0.719 was seen. Finally, hybrid models were developed by

combining various features and achieved maximum accuracy of 0.799 with

sensitivity 0.781 and specificity 0.818. Moreover, the performance of this model

was tested on an independent dataset that achieved an accuracy of 0.80. In

addition, we also compared the performance of our method with various

siRNA-designing methods on miRNA and siRNA datasets.

CONCLUSION: In this study, first time a method has been developed to predict

guide miRNA strands, of miRNA duplex. This study demonstrates that guide and

passenger strand of miRNA precursors can be distinguished using their nucleotide

sequence and secondary structure. This method will be useful in understanding

microRNA processing and can be implemented in RNA silencing technology to improve

the biological and clinical research. A web server has been developed based on

SVM models described in this study (http://crdd.osdd.net:8081/RISCbinder/).

DOI: 10.1186/1471-2105-10-105

PMCID: PMC2676257

PMID: 19358699 [Indexed for MEDLINE]

2271. Acta Biochim Biophys Sin (Shanghai). 2009 Apr;41(4):273-9.

A platform to standardize, store, and visualize proteomics experimental data.

Zheng G(1), Li H, Wang C, Sheng Q, Fan H, Yang S, Liu B, Dai J, Zeng R, Xie L.

Author information:

(1)Bioinformation Center, Key Laboratory of Systems Biology, Shanghai Institutes

for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China.

With the development of functional genomics research, large-scale proteomics

studies are now widespread, presenting significant challenges for data storage,

exchange, and analysis. Here we present the Integrated Proteomics Exploring

Database (IPED) as a platform for managing proteomics experimental data (both

process and result data). IPED is based on the schema of the Proteome

Experimental Data Repository (PEDRo), and complies with the General Proteomics

Standard (GPS) drafted by the Proteomics Standards Committee of the Human

Proteome Organization. In our work, we developed three components for the IPED

platform: the IPED client editor, IPED server software, and IPED web interface.

The client editor collects experimental data and generates an extensible markup

language (XML) data file compliant with PEDRo and GPS; the server software parses

the XML data file and loads information into a core database; and the web

interface displays experimental results, to provide a convenient graphic

representation of data. Given software convenience and data abundance, IPED is a

powerful platform for data exchange and presents an important resource for the

proteomics community. In its current release, IPED is available at

http://www.biosino.org/iped2.

PMID: 19352541 [Indexed for MEDLINE]

2272. Bioinformatics. 2009 Apr 1;25(7):977-8. doi: 10.1093/bioinformatics/btp081. Epub

2009 Feb 17.

OnTheFly: a tool for automated document-based text annotation, data linking and

network generation.

Pavlopoulos GA(1), Pafilis E, Kuhn M, Hooper SD, Schneider R.

Author information:

(1)Structural and Computational Biology Unit, EMBL Meyerhofstrasse 1, Heidelberg,

. pavlopou@embl.de

OnTheFly is a web-based application that applies biological named entity

recognition to enrich Microsoft Office, PDF and plain text documents. The input

files are converted into the HTML format and then sent to the Reflect tagging

server, which highlights biological entity names like genes, proteins and

chemicals, and attaches to them JavaScript code to invoke a summary pop-up

window. The window provides an overview of relevant information about the entity,

such as a protein description, the domain composition, a link to the 3D structure

and links to other relevant online resources. OnTheFly is also able to extract

the bioentities mentioned in a set of files and to produce a graphical

representation of the networks of the known and predicted associations of these

entities by retrieving the information from the STITCH database.AVAILABILITY:

http://onthefly.embl.de, http://onthefly.embl.de/FAQ.html.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp081

PMCID: PMC2660876

PMID: 19223449 [Indexed for MEDLINE]

2273. Bioinformatics. 2009 Apr 1;25(7):958-9. doi: 10.1093/bioinformatics/btp086. Epub

2009 Feb 13.

CGAS: comparative genomic analysis server.

Itoh M(1), Watanabe H.

Author information:

(1)Graduate School of Information Science and Technology, Hokkaido University,

Sapporo, Hokkaido, Japan.

SUMMARY: Comparative approach is one of the most essential methods for extracting

functional and evolutionary information from genomic sequences. So far, a number

of sequence comparison tools have been developed, and most are either for on-site

use, requiring program installation but providing a wide variety of analyses, or

for the online search of user's sequences against given databases on a server. We

newly devised an Asynchronous JavaScript and XML (Ajax)-based system for

comparative genomic analyses, CGAS, with highly interactive interface within a

browser, requiring no software installation. The current version, CGAS version 1,

provides functionality for viewing similarity relationships between user's

sequences, including a multiple dot plot between sequences with their annotation

information. The scrollbar-less 'draggable' interface of CGAS is implemented with

Google Maps API version 2. The annotation information associated with the genomic

sequences compared is synchronously displayed with the comparison view. The

multiple-comparison viewer is one of the unique functionalities of this system to

allow the users to compare the differences between different pairs of sequences.

In this viewer, the system tells orthologous correspondences between the

sequences compared interactively. This web-based tool is platform-independent and

will provide biologists having no computational skills with opportunities to

analyze their own data without software installation and customization of the

computer system.

AVAILABILITY AND IMPLEMENTATION: CGAS is available at

http://cgas.ist.hokudai.ac.jp/.

DOI: 10.1093/bioinformatics/btp086

PMCID: PMC2660877

PMID: 19218352 [Indexed for MEDLINE]

2274. Bioinformatics. 2009 Apr 1;25(7):964-6. doi: 10.1093/bioinformatics/btp021. Epub

2009 Feb 4.

QVALITY: non-parametric estimation of q-values and posterior error probabilities.

Käll L(1), Storey JD, Noble WS.

Author information:

(1)Department of Biochemistry and Biophysics, Center for Biomembrane Research,

Stockholm University, Sweden, USA. lukas.kall@cbr.su.se

Qvality is a C++ program for estimating two types of standard statistical

confidence measures: the q-value, which is an analog of the p-value that

incorporates multiple testing correction, and the posterior error probability

(PEP, also known as the local false discovery rate), which corresponds to the

probability that a given observation is drawn from the null distribution. In

computing q-values, qvality employs a standard bootstrap procedure to estimate

the prior probability of a score being from the null distribution; for PEP

estimation, qvality relies upon non-parametric logistic regression. Relative to

other tools for estimating statistical confidence measures, qvality is unique in

its ability to estimate both types of scores directly from a null distribution,

without requiring the user to calculate p-values.AVAILABILITY: A web server, C++

source code and binaries are available under MIT license at

http://noble.gs.washington.edu/proj/qvality.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp021

PMCID: PMC2660870

PMID: 19193729 [Indexed for MEDLINE]

2275. Comput Methods Programs Biomed. 2009 Apr;94(1):26-38. doi:

10.1016/j.cmpb.2008.10.004. Epub 2008 Nov 26.

Development of a web database portfolio system with PACS connectivity for

undergraduate health education and continuing professional development.

Ng CK(1), White P, McKay JC.

Author information:

(1)Department of Health Technology & Informatics, The Hong Kong Polytechnic

University, Hung Hom, Kowloon, Hong Kong. curtise ng@hotmail.com

Increasingly, the use of web database portfolio systems is noted in medical and

health education, and for continuing professional development (CPD). However, the

functions of existing systems are not always aligned with the corresponding

pedagogy and hence reflection is often lost. This paper presents the development

of a tailored web database portfolio system with Picture Archiving and

Communication System (PACS) connectivity, which is based on the portfolio

pedagogy. Following a pre-determined portfolio framework, a system model with the

components of web, database and mail servers, server side scripts, and a

Query/Retrieve (Q/R) broker for conversion between Hypertext Transfer Protocol

(HTTP) requests and Q/R service class of Digital Imaging and Communication in

Medicine (DICOM) standard, is proposed. The system was piloted with seventy-seven

volunteers. A tailored web database portfolio system

(http://radep.hti.polyu.edu.hk) was developed. Technological arrangements for

reinforcing portfolio pedagogy include popup windows (reminders) with guidelines

and probing questions of 'collect', 'select' and 'reflect' on evidence of

development/experience, limitation in the number of files (evidence) to be

uploaded, the 'Evidence Insertion' functionality to link the individual uploaded

artifacts with reflective writing, capability to accommodate diversity of

contents and convenient interfaces for reviewing portfolios and communication.

Evidence to date suggests the system supports users to build their portfolios

with sound hypertext reflection under a facilitator's guidance, and with

reviewers to monitor students' progress providing feedback and comments online in

a programme-wide situation.

DOI: 10.1016/j.cmpb.2008.10.004

PMID: 19038474 [Indexed for MEDLINE]

2276. PLoS Comput Biol. 2009 Apr;5(4):e1000352. doi: 10.1371/journal.pcbi.1000352. Epub

2009 Apr 10.

Statistical methods for detecting differentially abundant features in clinical

metagenomic samples.

White JR(1), Nagarajan N, Pop M.

Author information:

(1)Applied Mathematics and Scientific Computation Program, Center for

Bioinformatics and Computational Biology, University of Maryland, College Park,

Maryland, United States of America.

Numerous studies are currently underway to characterize the microbial communities

inhabiting our world. These studies aim to dramatically expand our understanding

of the microbial biosphere and, more importantly, hope to reveal the secrets of

the complex symbiotic relationship between us and our commensal bacterial

microflora. An important prerequisite for such discoveries are computational

tools that are able to rapidly and accurately compare large datasets generated

from complex bacterial communities to identify features that distinguish them.We

present a statistical method for comparing clinical metagenomic samples from two

treatment populations on the basis of count data (e.g. as obtained through

sequencing) to detect differentially abundant features. Our method, Metastats,

employs the false discovery rate to improve specificity in high-complexity

environments, and separately handles sparsely-sampled features using Fisher's

exact test. Under a variety of simulations, we show that Metastats performs well

compared to previously used methods, and significantly outperforms other methods

for features with sparse counts. We demonstrate the utility of our method on

several datasets including a 16S rRNA survey of obese and lean human gut

microbiomes, COG functional profiles of infant and mature gut microbiomes, and

bacterial and viral metabolic subsystem data inferred from random sequencing of

85 metagenomes. The application of our method to the obesity dataset reveals

differences between obese and lean subjects not reported in the original study.

For the COG and subsystem datasets, we provide the first statistically rigorous

assessment of the differences between these populations. The methods described in

this paper are the first to address clinical metagenomic datasets comprising

samples from multiple subjects. Our methods are robust across datasets of varied

complexity and sampling level. While designed for metagenomic applications, our

software can also be applied to digital gene expression studies (e.g. SAGE). A

web server implementation of our methods and freely available source code can be

found at http://metastats.cbcb.umd.edu/.

DOI: 10.1371/journal.pcbi.1000352

PMCID: PMC2661018

PMID: 19360128 [Indexed for MEDLINE]

2277. Virchows Arch. 2009 Apr;454(4):421-9. doi: 10.1007/s00428-009-0749-3. Epub 2009

Mar 12.

A European network for virtual microscopy--design, implementation and evaluation

of performance.

Lundin M(1), Szymas J, Linder E, Beck H, de Wilde P, van Krieken H, García Rojo

M, Moreno I, Ariza A, Tuzlali S, Dervişoğlu S, Helin H, Lehto VP, Lundin J.

Author information:

(1)Department of Oncology, Institute of Clinical Medicine, University of

Helsinki, Helsinki, Finland. mikael.lundin@helsinki.fi

Web-based virtual microscopy has enabled new applications within pathology. Here,

we introduce and evaluate a network of academic servers, designed to maximize

image accessibility to users from all regions of Europe. Whole-slide imaging was

utilized to digitize the entire slide set (n = 154) for the slide seminars of the

21st European Congress of Pathology. The virtual slides were mirrored to five

academic servers across Europe using a novel propagation method. Functionality

was implemented that automatically selects the fastest server connection in order

to optimize the slide-viewing speed ( http://www.webmicroscope.net/ECP2007).

Results show that during 6 months of monitoring the uptime of the network was

100%. The average viewing speed with the network was 3.1 Mbit/s, as compared to

1.9 Mbit/s using single servers. A good viewing speed (>2Mbit/s) was observed in

32 of 37 countries (86%), compared to 25 of 37 (68%) using single servers. Our

study shows that implementing a virtual microscopy network spanning a large

geographical area is technically feasible. By utilizing existing academic

networks and cost-minimizing image compression, it is also economically feasible.

DOI: 10.1007/s00428-009-0749-3

PMID: 19280223 [Indexed for MEDLINE]

2278. Biophys J. 2009 Mar 18;96(6):2119-27. doi: 10.1016/j.bpj.2008.12.3898.

Protein structure prediction by pro-Sp3-TASSER.

Zhou H(1), Skolnick J.

Author information:

(1)Center for the Study of Systems Biology, School of Biology, Georgia Institute

of Technology, Atlanta, Georgia, USA.

An automated protein structure prediction algorithm,

pro-sp3-Threading/ASSEmbly/Refinement (TASSER), is described and benchmarked.

Structural templates are identified using five different scoring functions

derived from the previously developed threading methods PROSPECTOR\_3 and SP(3).

Top templates identified by each scoring function are combined to derive contact

and distant restraints for subsequent model refinement by short TASSER

simulations. For Medium/Hard targets (those with moderate to poor quality

templates and/or alignments), alternative template alignments are also generated

by parametric alignment and the top models selected by TASSER-QA are included in

the contact and distance restraint derivation. Then, multiple short TASSER

simulations are used to generate an ensemble of full-length models. Subsequently,

the top models are selected from the ensemble by TASSER-QA and used to derive

TASSER contacts and distant restraints for another round of full TASSER

refinement. The final models are selected from both rounds of TASSER simulations

by TASSER-QA. We compare pro-sp3-TASSER with our previously developed MetaTASSER

method (enhanced with chunk-TASSER for Medium/Hard targets) on a representative

test data set of 723 proteins <250 residues in length. For the 348 proteins

classified as easy targets (those templates with good alignments and global

structure similarity to the target), the cumulative TM-score of the best of top

five models by pro-sp3-TASSER shows a 2.1% improvement over MetaTASSER. For the

155/220 medium/hard targets, the improvements in TM-score are 2.8% and 2.2%,

respectively. All improvements are statistically significant. More importantly,

the number of foldable targets (those having models whose TM-score to native >0.4

in the top five clusters) increases from 472 to 497 for all targets, and the

relative increases for medium and hard targets are 10% and 15%, respectively. A

server that implements the above algorithm is available at

http://cssb.biology.gatech.edu/skolnick/webservice/pro-sp3-TASSER/. The source

code is also available upon request.

DOI: 10.1016/j.bpj.2008.12.3898

PMCID: PMC2717286

PMID: 19289038 [Indexed for MEDLINE]

2279. Bioinformatics. 2009 Mar 15;25(6):838-40. doi: 10.1093/bioinformatics/btp049.

Epub 2009 Feb 2.

SciMiner: web-based literature mining tool for target identification and

functional enrichment analysis.

Hur J(1), Schuyler AD, States DJ, Feldman EL.

Author information:

(1)Bioinformatics Program, University of Michigan, Ann Arbor, MI 48109, USA.

juhur@umich.edu

SciMiner is a web-based literature mining and functional analysis tool that

identifies genes and proteins using a context specific analysis of MEDLINE

abstracts and full texts. SciMiner accepts a free text query (PubMed Entrez

search) or a list of PubMed identifiers as input. SciMiner uses both regular

expression patterns and dictionaries of gene symbols and names compiled from

multiple sources. Ambiguous acronyms are resolved by a scoring scheme based on

the co-occurrence of acronyms and corresponding description terms, which

incorporates optional user-defined filters. Functional enrichment analyses are

used to identify highly relevant targets (genes and proteins), GO (Gene Ontology)

terms, MeSH (Medical Subject Headings) terms, pathways and protein-protein

interaction networks by comparing identified targets from one search result with

those from other searches or to the full HGNC [HUGO (Human Genome Organization)

Gene Nomenclature Committee] gene set. The performance of gene/protein name

identification was evaluated using the BioCreAtIvE (Critical Assessment of

Information Extraction systems in Biology) version 2 (Year 2006) Gene

Normalization Task as a gold standard. SciMiner achieved 87.1% recall, 71.3%

precision and 75.8% F-measure. SciMiner's literature mining performance coupled

with functional enrichment analyses provides an efficient platform for retrieval

and summary of rich biological information from corpora of users'

interests.AVAILABILITY: http://jdrf.neurology.med.umich.edu/SciMiner/. A server

version of the SciMiner is also available for download and enables users to

utilize their institution's journal subscriptions.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btp049

PMCID: PMC2654801

PMID: 19188191 [Indexed for MEDLINE]

2280. Bioinformatics. 2009 Mar 15;25(6):830-1. doi: 10.1093/bioinformatics/btp055. Epub

2009 Jan 28.

The DICS repository: module-assisted analysis of disease-related gene lists.

Dietmann S(1), Georgii E, Antonov A, Tsuda K, Mewes HW.

Author information:

(1)Institute for Bioinformatics and Systems Biology, German Research Center for

Environmental Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany.

sabine.dietmann@googlemail.com

SUMMARY: The DICS database is a dynamic web repository of computationally

predicted functional modules from the human protein-protein interaction network.

It provides references to the CORUM, DrugBank, KEGG and Reactome pathway

databases. DICS can be accessed for retrieving sets of overlapping modules and

protein complexes that are significantly enriched in a gene list, thereby

providing valuable information about the functional context.

AVAILABILITY: Supplementary information on datasets and methods is available on

the web server http://mips.gsf.de/proj/dics.

DOI: 10.1093/bioinformatics/btp055

PMID: 19176557 [Indexed for MEDLINE]

2281. Bioinformatics. 2009 Mar 1;25(5):621-7. doi: 10.1093/bioinformatics/btp036. Epub

2009 Jan 28.

Fragment-based identification of druggable 'hot spots' of proteins using Fourier

domain correlation techniques.

Brenke R(1), Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda

S.

Author information:

(1)Program in Bioinformatics, Boston University, Boston, MA, USA.

MOTIVATION: The binding sites of proteins generally contain smaller regions that

provide major contributions to the binding free energy and hence are the prime

targets in drug design. Screening libraries of fragment-sized compounds by NMR or

X-ray crystallography demonstrates that such 'hot spot' regions bind a large

variety of small organic molecules, and that a relatively high 'hit rate' is

predictive of target sites that are likely to bind drug-like ligands with high

affinity. Our goal is to determine the 'hot spots' computationally rather than

experimentally.

RESULTS: We have developed the FTMAP algorithm that performs global search of the

entire protein surface for regions that bind a number of small organic probe

molecules. The search is based on the extremely efficient fast Fourier transform

(FFT) correlation approach which can sample billions of probe positions on dense

translational and rotational grids, but can use only sums of correlation

functions for scoring and hence is generally restricted to very simple energy

expressions. The novelty of FTMAP is that we were able to incorporate and

represent on grids a detailed energy expression, resulting in a very accurate

identification of low-energy probe clusters. Overlapping clusters of different

probes are defined as consensus sites (CSs). We show that the largest CS is

generally located at the most important subsite of the protein binding site, and

the nearby smaller CSs identify other important subsites. Mapping results are

presented for elastase whose structure has been solved in aqueous solutions of

eight organic solvents, and we show that FTMAP provides very similar information.

The second application is to renin, a long-standing pharmaceutical target for the

treatment of hypertension, and we show that the major CSs trace out the shape of

the first approved renin inhibitor, aliskiren.

AVAILABILITY: FTMAP is available as a server at http://ftmap.bu.edu/.

DOI: 10.1093/bioinformatics/btp036

PMCID: PMC2647826

PMID: 19176554 [Indexed for MEDLINE]

2282. Bioinformatics. 2009 Mar 1;25(5):670-1. doi: 10.1093/bioinformatics/btp024. Epub

2009 Jan 19.

FrameDP: sensitive peptide detection on noisy matured sequences.

Gouzy J(1), Carrere S, Schiex T.

Author information:

(1)Laboratoire Interactions Plantes Micro-organismes (LIPM) UMR441/2594,

INRA/CNRS, F-31320 Castanet Tolosan, France.

SUMMARY: Transcriptome sequencing represents a fundamental source of information

for genome-wide studies and transcriptome analysis and will become increasingly

important for expression analysis as new sequencing technologies takes over array

technology. The identification of the protein-coding region in transcript

sequences is a prerequisite for systematic amino acid-level analysis and more

specifically for domain identification. In this article, we present FrameDP, a

self-training integrative pipeline for predicting CDS in transcripts which can

adapt itself to different levels of sequence qualities.

AVAILABILITY: FrameDP for Linux (web-server and underlying pipeline) is available

at {{http://iant.toulouse.inra.fr/FrameDP}} for direct use or a standalone

installation.

DOI: 10.1093/bioinformatics/btp024

PMCID: PMC2647831

PMID: 19153134 [Indexed for MEDLINE]

2283. Bioinformatics. 2009 Mar 1;25(5):676-7. doi: 10.1093/bioinformatics/btp034. Epub

2009 Jan 16.

CPSP-web-tools: a server for 3D lattice protein studies.

Mann M(1), Smith C, Rabbath M, Edwards M, Will S, Backofen R.

Author information:

(1)Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, 79016

Freiburg, Germany. cpsp@informatik.uni-freiburg.de

Studies on proteins are often restricted to highly simplified models to face the

immense computational complexity of the associated problems. Constraint-based

protein structure prediction (CPSP) tools is a package of very fast algorithms

for ab initio optimal structure prediction and related problems in 3D HP-models

[cubic and face centered cubic (FCC)]. Here, we present CPSP-web-tools, an

interactive online interface of these programs for their immediate use. They

include the first method for the direct prediction of optimal energies and

structures in 3D HP side-chain models. This newest extension of the CPSP approach

is described here for the first time.AVAILABILITY AND IMPLEMENTATION: Free access

at http://cpsp.informatik.uni-freiburg.de

DOI: 10.1093/bioinformatics/btp034

PMCID: PMC2647832

PMID: 19151096 [Indexed for MEDLINE]

2284. Bioinformatics. 2009 Mar 1;25(5):606-14. doi: 10.1093/bioinformatics/btp023. Epub

2009 Jan 15.

Principal component analysis of native ensembles of biomolecular structures

(PCA\_NEST): insights into functional dynamics.

Yang LW(1), Eyal E, Bahar I, Kitao A.

Author information:

(1)University of Tokyo, Institute of Molecular and Cellular Bioscience, Tokyo,

Japan. lwy1@iam.u-tokyo.ac.jp

Erratum in

Bioinformatics. 2009 Aug 15;25(16):2147.

MOTIVATION: To efficiently analyze the 'native ensemble of conformations'

accessible to proteins near their folded state and to extract essential

information from observed distributions of conformations, reliable mathematical

methods and computational tools are needed.

RESULT: Examination of 24 pairs of structures determined by both NMR and X-ray

reveals that the differences in the dynamics of the same protein resolved by the

two techniques can be tracked to the most robust low frequency modes elucidated

by principal component analysis (PCA) of NMR models. The active sites of enzymes

are found to be highly constrained in these PCA modes. Furthermore, the residues

predicted to be highly immobile are shown to be evolutionarily conserved, lending

support to a PCA-based identification of potential functional sites. An online

tool, PCA\_NEST, is designed to derive the principal modes of conformational

changes from structural ensembles resolved by experiments or generated by

computations.

AVAILABILITY: http://ignm.ccbb.pitt.edu/oPCA\_Online.htm

DOI: 10.1093/bioinformatics/btp023

PMCID: PMC2647834

PMID: 19147661 [Indexed for MEDLINE]

2285. Biotechnol Prog. 2009 Mar-Apr;25(2):409-16. doi: 10.1002/btpr.147.

Aspergillus niger lipase: Heterologous expression in Pichia pastoris, molecular

modeling prediction and the importance of the hinge domains at both sides of the

lid domain to interfacial activation.

Shu Z(1), Duan M, Yang J, Xu L, Yan Y.

Author information:

(1)Engineering Research Center of Industrial Microbiology, Ministry of Education,

Fujian Normal University, Fuzhou Fujian 350108, PR China. shuzhengyu@gmail.com

Aspergillus niger lipase (ANL) is an important biocatalyst in the food processing

industry. However, there is no report of its detailed three-dimensional structure

because of difficulties in crystallization. In this article, based on

experimental data and bioinformational analysis results, the structural features

of ANL were simulated. Firstly, two recombinant ANLs expressed in Pichia pastoris

were purified to homogeneity and their corresponding secondary structure

compositions were determined by circular dichroism spectra. Secondly, the primary

structure, the secondary structure and the three-dimensional structure of ANL

were modeled by comparison with homologous lipases with known three-dimensional

structures using the BioEdit software, lipase engineering database

(http://www.led.uni-stuttgart.de/), PSIPRED server and SwissModel server. The

predicted molecular structure of ANL presented typical features of the alpha/beta

hydrolase fold including positioning of the putative catalytic triad residues and

the GXSXG signature motif. Comparison of the predicted three-dimensional

structure of ANL with the X-ray three-dimensional structure of A. niger feruloyl

esterase showed that the functional difference of interfacial activation between

lipase and esterase was concerned with the difference in position of the lid. Our

three-dimensional model of ANL helps to modify lipase structure by protein

engineering, which will further expand the scope of application of ANL.

DOI: 10.1002/btpr.147

PMID: 19248178 [Indexed for MEDLINE]

2286. J Proteome Res. 2009 Mar;8(3):1577-84. doi: 10.1021/pr800957q.

QuatIdent: a web server for identifying protein quaternary structural attribute

by fusing functional domain and sequential evolution information.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong

University, 800 Dongchuan Road, Shanghai, 200240, China.

Many proteins exist in vivo as oligomers with various different quaternary

structural attributes rather than as single individual chains. They are the

structural bases of various marvelous biological functions such as cooperative

effects, allosteric mechanism, and ion-channel gating. Therefore, with the

avalanche of protein sequences generated in the postgenomic era, it is very

important for both basic research and drug discovery to identify their quaternary

structural attributes in a timely manner. In view of this, a powerful ensemble

identifier, called QuatIdent, is developed by fusing the functional domain and

sequential evolution information. QuatIdent is a 2-layer predictor. The 1st layer

is for identifying a query protein as belonging to which one of the following 10

main quaternary structural attributes: (1) monomer, (2) dimer, (3) trimer, (4)

tetramer, (5) pentamer, (6) hexamer, (7) heptamer, (8) octamer, (9) decamer, and

(10) dodecamer. If the result thus obtained turns out to be anything but monomer,

the process will be automatically continued to further identify it as belonging

to a homo-oligomer or hetero-oligomer. The overall success rate by QuatIdent for

the 1st layer identification was 71.1% and that for the 2nd layer ranged from 84

to 96%. These rates were derived by the jackknife cross-validation tests on the

stringent benchmark data sets where none of proteins has > or =60% pairwise

sequence identity to any other in a same subset. QuatIdent is freely accessible

to the public as a web server via the site at

http://www.csbio.sjtu.edu.cn/bioinf/Quaternary/ , by which one can get the

desired 2-level results for a query protein sequence in around 25 seconds. The

longer the sequence is, the more time that is needed.

DOI: 10.1021/pr800957q

PMID: 19226167 [Indexed for MEDLINE]

2287. PLoS Comput Biol. 2009 Mar;5(3):e1000304. doi: 10.1371/journal.pcbi.1000304. Epub

2009 Mar 13.

Detection of alpha-rod protein repeats using a neural network and application to

huntingtin.

Palidwor GA(1), Shcherbinin S, Huska MR, Rasko T, Stelzl U, Arumughan A, Foulle

R, Porras P, Sanchez-Pulido L, Wanker EE, Andrade-Navarro MA.

Author information:

(1)Ottawa Health Research Institute, Ottawa, Ontario, Canada.

A growing number of solved protein structures display an elongated structural

domain, denoted here as alpha-rod, composed of stacked pairs of anti-parallel

alpha-helices. Alpha-rods are flexible and expose a large surface, which makes

them suitable for protein interaction. Although most likely originating by tandem

duplication of a two-helix unit, their detection using sequence similarity

between repeats is poor. Here, we show that alpha-rod repeats can be detected

using a neural network. The network detects more repeats than are identified by

domain databases using multiple profiles, with a low level of false positives

(<10%). We identify alpha-rod repeats in approximately 0.4% of proteins in

eukaryotic genomes. We then investigate the results for all human proteins,

identifying alpha-rod repeats for the first time in six protein families,

including proteins STAG1-3, SERAC1, and PSMD1-2 & 5. We also characterize a short

version of these repeats in eight protein families of Archaeal, Bacterial, and

Fungal species. Finally, we demonstrate the utility of these predictions in

directing experimental work to demarcate three alpha-rods in huntingtin, a

protein mutated in Huntington's disease. Using yeast two hybrid analysis and an

immunoprecipitation technique, we show that the huntingtin fragments containing

alpha-rods associate with each other. This is the first definition of domains in

huntingtin and the first validation of predicted interactions between fragments

of huntingtin, which sets up directions toward functional characterization of

this protein. An implementation of the repeat detection algorithm is available as

a Web server with a simple graphical output: http://www.ogic.ca/projects/ard.

This can be further visualized using BiasViz, a graphic tool for representation

of multiple sequence alignments.

DOI: 10.1371/journal.pcbi.1000304

PMCID: PMC2647740

PMID: 19282972 [Indexed for MEDLINE]

2288. PLoS Comput Biol. 2009 Mar;5(3):e1000307. doi: 10.1371/journal.pcbi.1000307. Epub

2009 Mar 13.

Probabilistic interaction network of evidence algorithm and its application to

complete labeling of peak lists from protein NMR spectroscopy.

Bahrami A(1), Assadi AH, Markley JL, Eghbalnia HR.

Author information:

(1)Biochemistry Department, National Magnetic Resonance Facility at Madison,

University of Wisconsin Madison, Madison, Wisconsin, United States of America.

arash@nmrfam.wisc.edu

The process of assigning a finite set of tags or labels to a collection of

observations, subject to side conditions, is notable for its computational

complexity. This labeling paradigm is of theoretical and practical relevance to a

wide range of biological applications, including the analysis of data from DNA

microarrays, metabolomics experiments, and biomolecular nuclear magnetic

resonance (NMR) spectroscopy. We present a novel algorithm, called Probabilistic

Interaction Network of Evidence (PINE), that achieves robust, unsupervised

probabilistic labeling of data. The computational core of PINE uses estimates of

evidence derived from empirical distributions of previously observed data, along

with consistency measures, to drive a fictitious system M with Hamiltonian H to a

quasi-stationary state that produces probabilistic label assignments for relevant

subsets of the data. We demonstrate the successful application of PINE to a key

task in protein NMR spectroscopy: that of converting peak lists extracted from

various NMR experiments into assignments associated with probabilities for their

correctness. This application, called PINE-NMR, is available from a freely

accessible computer server (http://pine.nmrfam.wisc.edu). The PINE-NMR server

accepts as input the sequence of the protein plus user-specified combinations of

data corresponding to an extensive list of NMR experiments; it provides as output

a probabilistic assignment of NMR signals (chemical shifts) to sequence-specific

backbone and aliphatic side chain atoms plus a probabilistic determination of the

protein secondary structure. PINE-NMR can accommodate prior information about

assignments or stable isotope labeling schemes. As part of the analysis, PINE-NMR

identifies, verifies, and rectifies problems related to chemical shift

referencing or erroneous input data. PINE-NMR achieves robust and consistent

results that have been shown to be effective in subsequent steps of NMR structure

determination.

DOI: 10.1371/journal.pcbi.1000307

PMCID: PMC2645676

PMID: 19282963 [Indexed for MEDLINE]

2289. Proteins. 2009 Mar;74(4):847-56. doi: 10.1002/prot.22193.

Improving the prediction accuracy of residue solvent accessibility and real-value

backbone torsion angles of proteins by guided-learning through a two-layer neural

network.

Faraggi E(1), Xue B, Zhou Y.

Author information:

(1)Indiana University School of Informatics, Indiana University-Purdue

University, Indianapolis, IN 46202, USA.

This article attempts to increase the prediction accuracy of residue solvent

accessibility and real-value backbone torsion angles of proteins through improved

learning. Most methods developed for improving the backpropagation algorithm of

artificial neural networks are limited to small neural networks. Here, we

introduce a guided-learning method suitable for networks of any size. The method

employs a part of the weights for guiding and the other part for training and

optimization. We demonstrate this technique by predicting residue solvent

accessibility and real-value backbone torsion angles of proteins. In this

application, the guiding factor is designed to satisfy the intuitive condition

that for most residues, the contribution of a residue to the structural

properties of another residue is smaller for greater separation in the

protein-sequence distance between the two residues. We show that the

guided-learning method makes a 2-4% reduction in 10-fold cross-validated mean

absolute errors (MAE) for predicting residue solvent accessibility and backbone

torsion angles, regardless of the size of database, the number of hidden layers

and the size of input windows. This together with introduction of two-layer

neural network with a bipolar activation function leads to a new method that has

a MAE of 0.11 for residue solvent accessibility, 36 degrees for psi, and 22

degrees for phi. The method is available as a Real-SPINE 3.0 server in

http://sparks.informatics.iupui.edu.

DOI: 10.1002/prot.22193

PMCID: PMC2635924

PMID: 18704931 [Indexed for MEDLINE]

2290. J Biomed Sci. 2009 Feb 24;16:25. doi: 10.1186/1423-0127-16-25.

An integrated method for cancer classification and rule extraction from

microarray data.

Huang LT(1).

Author information:

(1)Department of Computer Science and Information Engineering, Mingdao

University, Changhua 523, Taiwan. larry@mdu.edu.tw

Different microarray techniques recently have been successfully used to

investigate useful information for cancer diagnosis at the gene expression level

due to their ability to measure thousands of gene expression levels in a

massively parallel way. One important issue is to improve classification

performance of microarray data. However, it would be ideal that influential genes

and even interpretable rules can be explored at the same time to offer biological

insight. Introducing the concepts of system design in software engineering, this

paper has presented an integrated and effective method (named X-AI) for accurate

cancer classification and the acquisition of knowledge from DNA microarray data.

This method included a feature selector to systematically extract the relative

important genes so as to reduce the dimension and retain as much as possible of

the class discriminatory information. Next, diagonal quadratic discriminant

analysis (DQDA) was combined to classify tumors, and generalized rule induction

(GRI) was integrated to establish association rules which can give an

understanding of the relationships between cancer classes and related genes. Two

non-redundant datasets of acute leukemia were used to validate the proposed X-AI,

showing significantly high accuracy for discriminating different classes. On the

other hand, I have presented the abilities of X-AI to extract relevant genes, as

well as to develop interpretable rules. Further, a web server has been

established for cancer classification and it is freely available at

http://bioinformatics.myweb.hinet.net/xai.htm.

DOI: 10.1186/1423-0127-16-25

PMCID: PMC2653531

PMID: 19272192 [Indexed for MEDLINE]

2291. Bioinformatics. 2009 Feb 15;25(4):543-4. doi: 10.1093/bioinformatics/btp008. Epub

2009 Jan 6.

BioCichlid: central dogma-based 3D visualization system of time-course microarray

data on a hierarchical biological network.

Ishiwata RR(1), Morioka MS, Ogishima S, Tanaka H.

Author information:

(1)Department of Computational Biology, School of Medical Sciences, Tokyo Medical

and Dental University, Yushima 1-5-45, Bunkyo-ku, Tokyo 113-8510, Japan.

SUMMARY: BioCichlid is a 3D visualization system of time-course microarray data

on molecular networks, aiming at interpretation of gene expression data by

transcriptional relationships based on the central dogma with physical and

genetic interactions. BioCichlid visualizes both physical (protein) and genetic

(regulatory) network layers, and provides animation of time-course gene

expression data on the genetic network layer. Transcriptional regulations are

represented to bridge the physical network (transcription factors) and genetic

network (regulated genes) layers, thus integrating promoter analysis into the

pathway mapping. BioCichlid enhances the interpretation of microarray data and

allows for revealing the underlying mechanisms causing differential gene

expressions.

AVAILABILITY: BioCichlid is freely available and can be accessed at

http://newton.tmd.ac.jp/. Source codes for both biocichlid server and client are

also available.

DOI: 10.1093/bioinformatics/btp008

PMID: 19126577 [Indexed for MEDLINE]

2292. J Theor Biol. 2009 Feb 7;256(3):441-6. doi: 10.1016/j.jtbi.2008.10.007. Epub 2008

Oct 19.

Predicting protein fold pattern with functional domain and sequential evolution

information.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong

University, 800 Dongchuan Road, Shanghai 200240, China. hbshen@sjtu.edu.cn

The fold pattern of a protein is one level deeper than its structural

classification, and hence is more challenging and complicated for prediction.

Many efforts have been made in this regard, but so far all the reported success

rates are still under 70%, indicating that it is extremely difficult to enhance

the success rate even by 1% or 2%. To address this problem, here a novel approach

is proposed that is featured by combining the functional domain information and

the sequential evolution information through a fusion ensemble classifier. The

predictor thus developed is called PFP-FunDSeqE. Tests were performed for

identifying proteins among their 27 fold patterns. Compared with the existing

predictors tested by a same stringent benchmark dataset, the new predictor can,

for the first time, achieve over 70% success rate. The PFP-FunDSeqE predictor is

freely available to the public as a web server at

http://www.csbio.sjtu.edu.cn/bioinf/PFP-FunDSeqE/.

DOI: 10.1016/j.jtbi.2008.10.007

PMID: 18996396 [Indexed for MEDLINE]

2293. Anal Biochem. 2009 Feb 1;385(1):153-60. doi: 10.1016/j.ab.2008.10.020. Epub 2008

Nov 1.

Identification of proteases and their types.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, Shanghai 200240, China. hbshen@sjtu.edu.cn

Called by many as biology's version of Swiss army knives, proteases cut long

sequences of amino acids into fragments and regulate most physiological

processes. They are vitally important in the life cycle. Different types of

proteases have different action mechanisms and biological processes. With the

avalanche of protein sequences generated during the postgenomic age, it is highly

desirable for both basic research and drug design to develop a fast and reliable

method for identifying the types of proteases according to their sequences or

even just for whether they are proteases or not. In this article, three recently

developed identification methods in this regard are discussed: (i) FunD-PseAAC,

(ii) GO-PseAAC, and (iii) FunD-PsePSSM. The first two were established by

hybridizing the FunD (functional domain) approach and the GO (gene ontology)

approach, respectively, with the PseAAC (pseudo amino acid composition) approach.

The third method was established by fusing the FunD approach with the PsePSSM

(pseudo position-specific scoring matrix) approach. Of these three methods, only

FunD-PsePSSM has provided a server called ProtIdent (protease identifier), which

is freely accessible to the public via the website at

http://www.csbio.sjtu.edu.cn/bioinf/Protease. For the convenience of users, a

step-by-step guide on how to use ProtIdent is illustrated. Meanwhile, the caveat

in using ProtIdent and how to understand the success expectancy rate of a

statistical predictor are discussed. Finally, the essence of why ProtIdent can

yield a high success rate in identifying proteases and their types is elucidated.

DOI: 10.1016/j.ab.2008.10.020

PMID: 19007742 [Indexed for MEDLINE]

2294. Assay Drug Dev Technol. 2009 Feb;7(1):44-55. doi: 10.1089/adt.2008.174.

WebFlow: a software package for high-throughput analysis of flow cytometry data.

Hammer MM(1), Kotecha N, Irish JM, Nolan GP, Krutzik PO.

Author information:

(1)Baxter Laboratory in Genetic Pharmacology, Department of Microbiology,

Stanford University, Stanford, California 94305, USA.

Flow cytometry has emerged as a powerful tool for quantitative, single-cell

analysis of both surface markers and intracellular antigens, including

phosphoproteins and kinase signaling cascades, with the flexibility to process

hundreds of samples in multiwell plate format. Quantitative flow cytometric

analysis is being applied in many areas of biology, from the study of immunology

in animal models or human patients to high-content drug screening of

pharmacologically active compounds. However, these experiments generate thousands

of data points per sample, each with multiple measured parameters, leading to

data management and analysis challenges. We developed WebFlow

(http://webflow.stanford.edu), a web server-based software package to manage,

analyze, and visualize data from flow cytometry experiments. WebFlow is

accessible via standard web browsers and does not require users to install

software on their personal computers. The software enables plate-based annotation

of large data sets, which provides the basis for exploratory data analysis tools

and rapid visualization of multiple different parameters. These tools include

custom user-defined statistics to normalize data to other wells or other

channels, as well as interactive, user-selectable heat maps for viewing the

underlying single-cell data. The web-based approach of WebFlow allows for sharing

of data with collaborators or the general public. WebFlow provides a novel

platform for quantitative analysis of flow cytometric data from high-throughput

drug screening or disease profiling experiments.

DOI: 10.1089/adt.2008.174

PMCID: PMC2956679

PMID: 19187010 [Indexed for MEDLINE]

2295. Bioinformatics. 2009 Feb 1;25(3):413-4. doi: 10.1093/bioinformatics/btn584. Epub

2008 Nov 11.

ProtorP: a protein-protein interaction analysis server.

Reynolds C(1), Damerell D, Jones S.

Author information:

(1)Department of Biochemistry, School of Life Sciences, John Maynard Smith

Building, University of Sussex, Falmer, Brighton BN1 9QG, UK.

SUMMARY: The PROTORP server analyses protein-protein associations in 3D

structures. The server calculates a series of physical and chemical parameters of

the protein interaction sites that contribute to the binding energy of the

association. These parameters include, size and shape, intermolecular bonding,

residue and atom composition and secondary structure contributions. The server is

flexible, in that it allows users to analyse individual protein associations or

large datasets of associations deposited in the PDB, or upload and analyse

proprietary files. The properties calculated can be compared with parameter

distributions for non-homologous datasets of different classes of protein

associations provided on the server website. The server provides an efficient way

of characterizing protein-protein associations of new or existing proteins, and a

means of putting these values in the context of previously observed associations.

AVAILABILITY: http://www.bioinformatics.sussex.ac.uk/protorp

DOI: 10.1093/bioinformatics/btn584

PMID: 19001476 [Indexed for MEDLINE]

2296. Environ Microbiol Rep. 2009 Feb;1(1):78-85.

Online program 'vipcal' for calculating lytic viral production and lysogenic

cells based on a viral reduction approach.

Luef B(1), Luef F, Peduzzi P.

Author information:

(1)Faculty of Life Sciences, Department of Freshwater Ecology, University of

Vienna, Vienna, Austria.

Assessing viral production (VP) requires robust methodological settings combined

with precise mathematical calculations. This contribution improves and

standardizes mathematical calculations of VP and the assessment of the proportion

of lysogenic cells in a sample. We present an online tool 'Viral Production

Calculator' (vipcal, http://www.univie.ac.at/nuhag-php/vipcal) that calculates

lytic production and the percentage of lysogenic cells based on data obtained

from a viral reduction approach (VRA). The main advantage of our method lies in

its universal applicability, even to different piecewise-linear curves. We

demonstrate the application of our tool for calculating lytic VP and the

proportion of lysogenic bacteria in an environmental sample. The program can also

be used to calculate different parameters for estimating virus-induced mortality,

including the percentage of lytically infected cells, lysis rate of bacteria,

percentage of bacterial production lysed, proportion of bacterial loss per day,

viral turnover time as well as dissolved organic carbon and nitrogen release.

vipcal helps avoid differences in the calculation of VP and diverse viral

parameters between studies and laboratories, which facilities interpretation of

results. This tool represents a methodological step forward that can help improve

our understanding of the role of viral activity in aquatic systems.

DOI: 10.1111/j.1758-2229.2008.00008.x

PMCID: PMC2999826

PMID: 21151811

2297. BMC Bioinformatics. 2009 Jan 19;10:22. doi: 10.1186/1471-2105-10-22.

Prediction of nuclear proteins using SVM and HMM models.

Kumar M(1), Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India.

BACKGROUND: The nucleus, a highly organized organelle, plays important role in

cellular homeostasis. The nuclear proteins are crucial for chromosomal

maintenance/segregation, gene expression, RNA processing/export, and many other

processes. Several methods have been developed for predicting the nuclear

proteins in the past. The aim of the present study is to develop a new method for

predicting nuclear proteins with higher accuracy.

RESULTS: All modules were trained and tested on a non-redundant dataset and

evaluated using five-fold cross-validation technique. Firstly, Support Vector

Machines (SVM) based modules have been developed using amino acid and dipeptide

compositions and achieved a Mathews correlation coefficient (MCC) of 0.59 and

0.61 respectively. Secondly, we have developed SVM modules using split amino acid

compositions (SAAC) and achieved the maximum MCC of 0.66. Thirdly, a hidden

Markov model (HMM) based module/profile was developed for searching exclusively

nuclear and non-nuclear domains in a protein. Finally, a hybrid module was

developed by combining SVM module and HMM profile and achieved a MCC of 0.87 with

an accuracy of 94.61%. This method performs better than the existing methods when

evaluated on blind/independent datasets. Our method estimated 31.51%, 21.89%,

26.31%, 25.72% and 24.95% of the proteins as nuclear proteins in Saccharomyces

cerevisiae, Caenorhabditis elegans, Drosophila melanogaster, mouse and human

proteomes respectively. Based on the above modules, we have developed a web

server NpPred for predicting nuclear proteins

http://www.imtech.res.in/raghava/nppred/.

CONCLUSION: This study describes a highly accurate method for predicting nuclear

proteins. SVM module has been developed for the first time using SAAC for

predicting nuclear proteins, where amino acid composition of N-terminus and the

remaining protein were computed separately. In addition, our study is a first

documentation where exclusively nuclear and non-nuclear domains have been

identified and used for predicting nuclear proteins. The performance of the

method improved further by combining both approaches together.

DOI: 10.1186/1471-2105-10-22

PMCID: PMC2632991

PMID: 19152693 [Indexed for MEDLINE]

2298. BMC Bioinformatics. 2009 Jan 7;10:8. doi: 10.1186/1471-2105-10-8.

Large-scale prediction of long disordered regions in proteins using random

forests.

Han P(1), Zhang X, Norton RS, Feng ZP.

Author information:

(1)The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria

3052, Australia. pengfei.han@rmit.edu.au

BACKGROUND: Many proteins contain disordered regions that lack fixed

three-dimensional (3D) structure under physiological conditions but have

important biological functions. Prediction of disordered regions in protein

sequences is important for understanding protein function and in high-throughput

determination of protein structures. Machine learning techniques, including

neural networks and support vector machines have been widely used in such

predictions. Predictors designed for long disordered regions are usually less

successful in predicting short disordered regions. Combining prediction of short

and long disordered regions will dramatically increase the complexity of the

prediction algorithm and make the predictor unsuitable for large-scale

applications. Efficient batch prediction of long disordered regions alone is of

greater interest in large-scale proteome studies.

RESULTS: A new algorithm, IUPforest-L, for predicting long disordered regions

using the random forest learning model is proposed in this paper. IUPforest-L is

based on the Moreau-Broto auto-correlation function of amino acid indices (AAIs)

and other physicochemical features of the primary sequences. In 10-fold cross

validation tests, IUPforest-L can achieve an area of 89.5% under the receiver

operating characteristic (ROC) curve. Compared with existing disorder predictors,

IUPforest-L has high prediction accuracy and is efficient for predicting long

disordered regions in large-scale proteomes.

CONCLUSION: The random forest model based on the auto-correlation functions of

the AAIs within a protein fragment and other physicochemical features could

effectively detect long disordered regions in proteins. A new predictor,

IUPforest-L, was developed to batch predict long disordered regions in proteins,

and the server can be accessed from

http://dmg.cs.rmit.edu.au/IUPforest/IUPforest-L.php.

DOI: 10.1186/1471-2105-10-8

PMCID: PMC2637845

PMID: 19128505 [Indexed for MEDLINE]

2299. Identifying Patients with Familial Cancer Syndromes.

Sijmons RH.

In: Riegert-Johnson DL, Boardman LA, Hefferon T, Roberts M, editors. Cancer

Syndromes [Internet]. Bethesda (MD): National Center for Biotechnology

Information (US); 2009-.

2010 Feb 27.

The identification of families with familial cancer syndromes would be relatively

easy if all the genes for these syndromes were known and all newly diagnosed

cancer patients could be screened using genetic testing. However, even if all of

the genes were known, there would be several ethical, technical, financial, and

other barriers to the introduction of mass genetic screening. Nevertheless, we

have seen slow movement in this direction, with the suggested screening of all

newly diagnosed patients with colorectal cancer for Lynch syndrome as an example

(1). Still, for many years to come, the identification of familial cancer

syndromes in the majority of families will rely on the alertness of clinicians

and on the families’ awareness of the importance of their personal and family

cancer history (2). Key elements in the diagnostic process are formulating a

genetic differential diagnosis based on the cancer history and other traits. Some

of the diagnostic clues may be easily observed by the clinician, whereas others

need to be actively searched for. This chapter reviews the tell-tale signs of

hereditary cancer and also presents an online tool, the Familial Cancer Database

(www.facd.info), which may assist the clinician in the diagnostic process.

PMID: 21249759

2300. Bioinformatics. 2009 Jan 1;25(1):141-3. doi: 10.1093/bioinformatics/btn590. Epub

2008 Nov 13.

CRONOS: the cross-reference navigation server.

Waegele B(1), Dunger-Kaltenbach I, Fobo G, Montrone C, Mewes HW, Ruepp A.

Author information:

(1)Institute for Bioinformatics and Systems Biology (MIPS), Helmholtz Zentrum

München - German Research Center for Environmental Health (GmbH), Ingolstädter

Landstrasse 1, D-85764 Neuherberg, Germany.

brigitte.waegele@helmholtz-muenchen.de

Cross-mapping of gene and protein identifiers between different databases is a

tedious and time-consuming task. To overcome this, we developed CRONOS, a

cross-reference server that contains entries from five mammalian organisms

presented by major gene and protein information resources. Sequence similarity

analysis of the mapped entries shows that the cross-references are highly

accurate. In total, up to 18 different identifier types can be used for

identification of cross-references. The quality of the mapping could be improved

substantially by exclusion of ambiguous gene and protein names which were

manually validated. Organism-specific lists of ambiguous terms, which are

valuable for a variety of bioinformatics applications like text mining are

available for download.AVAILABILITY: CRONOS is freely available to non-commercial

users at http://mips.gsf.de/genre/proj/cronos/index.html, web services are

available at http://mips.gsf.de/CronosWSService/CronosWS?wsdl.

DOI: 10.1093/bioinformatics/btn590

PMCID: PMC2638938

PMID: 19010804 [Indexed for MEDLINE]

2301. Bioinformatics. 2009 Jan 1;25(1):30-5. doi: 10.1093/bioinformatics/btn583. Epub

2008 Nov 12.

Prediction of DNA-binding residues in proteins from amino acid sequences using a

random forest model with a hybrid feature.

Wu J(1), Liu H, Duan X, Ding Y, Wu H, Bai Y, Sun X.

Author information:

(1)State Key Laboratory of Bioelectronics, School of Biological Science and

Medical Engineering, Southeast University, Nanjing 210096, P. R. China.

MOTIVATION: In this work, we aim to develop a computational approach for

predicting DNA-binding sites in proteins from amino acid sequences. To avoid

overfitting with this method, all available DNA-binding proteins from the Protein

Data Bank (PDB) are used to construct the models. The random forest (RF)

algorithm is used because it is fast and has robust performance for different

parameter values. A novel hybrid feature is presented which incorporates

evolutionary information of the amino acid sequence, secondary structure (SS)

information and orthogonal binary vector (OBV) information which reflects the

characteristics of 20 kinds of amino acids for two physical-chemical properties

(dipoles and volumes of the side chains). The numbers of binding and non-binding

residues in proteins are highly unbalanced, so a novel scheme is proposed to deal

with the problem of imbalanced datasets by downsizing the majority class.

RESULTS: The results show that the RF model achieves 91.41% overall accuracy with

Matthew's correlation coefficient of 0.70 and an area under the receiver

operating characteristic curve (AUC) of 0.913. To our knowledge, the RF method

using the hybrid feature is currently the computationally optimal approach for

predicting DNA-binding sites in proteins from amino acid sequences without using

three-dimensional (3D) structural information. We have demonstrated that the

prediction results are useful for understanding protein-DNA interactions.

AVAILABILITY: DBindR web server implementation is freely available at

http://www.cbi.seu.edu.cn/DBindR/DBindR.htm.

DOI: 10.1093/bioinformatics/btn583

PMCID: PMC2638931

PMID: 19008251 [Indexed for MEDLINE]

2302. Bioinformatics. 2009 Jan 1;25(1):22-9. doi: 10.1093/bioinformatics/btn580. Epub

2008 Nov 13.

Predicting DNA recognition by Cys2His2 zinc finger proteins.

Persikov AV(1), Osada R, Singh M.

Author information:

(1)Lewis-Sigler Institute for Integrative Genomics and Department of Computer

Science, Princeton University, Princeton, NJ 08544, USA.

MOTIVATION: Cys(2)His(2) zinc finger (ZF) proteins represent the largest class of

eukaryotic transcription factors. Their modular structure and well-conserved

protein-DNA interface allow the development of computational approaches for

predicting their DNA-binding preferences even when no binding sites are known for

a particular protein. The 'canonical model' for ZF protein-DNA interaction

consists of only four amino acid nucleotide contacts per zinc finger domain.

RESULTS: We present an approach for predicting ZF binding based on support vector

machines (SVMs). While most previous computational approaches have been based

solely on examples of known ZF protein-DNA interactions, ours additionally

incorporates information about protein-DNA pairs known to bind weakly or not at

all. Moreover, SVMs with a linear kernel can naturally incorporate constraints

about the relative binding affinities of protein-DNA pairs; this type of

information has not been used previously in predicting ZF protein-DNA binding.

Here, we build a high-quality literature-derived experimental database of ZF-DNA

binding examples and utilize it to test both linear and polynomial kernels for

predicting ZF protein-DNA binding on the basis of the canonical binding model.

The polynomial SVM outperforms previously published prediction procedures as well

as the linear SVM. This may indicate the presence of dependencies between

contacts in the canonical binding model and suggests that modification of the

underlying structural model may result in further improved performance in

predicting ZF protein-DNA binding. Overall, this work demonstrates that methods

incorporating information about non-binding and relative binding of protein-DNA

pairs have great potential for effective prediction of protein-DNA interactions.

AVAILABILITY: An online tool for predicting ZF DNA binding is available at

http://compbio.cs.princeton.edu/zf/.

DOI: 10.1093/bioinformatics/btn580

PMCID: PMC2638941

PMID: 19008249 [Indexed for MEDLINE]

2303. Bioinformatics. 2009 Jan 1;25(1):121-2. doi: 10.1093/bioinformatics/btn567. Epub

2008 Nov 6.

RANKPROP: a web server for protein remote homology detection.

Melvin I(1), Weston J, Leslie C, Noble WS.

Author information:

(1)NEC Laboratories of America, Princeton, NJ, USA. iain@nec-labs.com

We present a large-scale implementation of the Rankprop protein homology ranking

algorithm in the form of an openly accessible web server. We use the NRDB40

PSI-BLAST all-versus-all protein similarity network of 1.1 million proteins to

construct the graph for the Rankprop algorithm, whereas previously, results were

only reported for a database of 108 000 proteins. We also describe two

algorithmic improvements to the original algorithm, including propagation from

multiple homologs of the query and better normalization of ranking scores, that

lead to higher accuracy and to scores with a probabilistic

interpretation.AVAILABILITY: The Rankprop web server and source code are

available at http://rankprop.gs.washington.edu

DOI: 10.1093/bioinformatics/btn567

PMCID: PMC2638939

PMID: 18990723 [Indexed for MEDLINE]

2304. Bioinformation. 2009;3(7):296-8. Epub 2009 Feb 27.

RICHEST--a web server for richness estimation in biological data.

Durden C(1), Dong Q.

Author information:

(1)Center for Genomics and Bioinformatics, Indiana University, Bloomington,

Indiana, USA.

Richness is defined as the number of distinct species or classes in a sample or

population. Although richness estimation is an important practice, it requires

mathematical and computational methods that are challenging to understand and

implement. We have developed a web server, RICHness ESTimator (RICHEST), which

implements three non-parametric statistical methods for richness estimation. Its

user-friendly web interface allows users to analyze and compare their data

conveniently over the web.AVAILABILITY: A web server hosting RICHEST is

accessible at http://richest.cgb.indiana.edu/cgi-bin/index.cgi and the software

is freely available for local installations.

PMCID: PMC2655047

PMID: 19293995

2305. Bioinformation. 2009;3(7):293-5. Epub 2009 Feb 27.

ScanMoment: a web server for combinatorial analysis of basic residues in nucleic

acid binding sites.

Massey SE(1).

Author information:

(1)Department of Biology, University of Puerto Rico - Rio Piedras, PO Box 23360,

San Juan, Puerto Rico 00931. stevenemassey@gmail.com

ScanMoment is a webserver designed to identify the presence of the basic faced

alpha-helix (BFAH) motif in the nucleic acid binding sites of proteins. The

program calculates the 'Basic Moment', a parameter that quantitizes the

distribution of basic residues on the surface of an alpha-helix. A sliding window

is used to generate a plot displaying regions of the protein sequence that

possesses a high Basic Moment and hus likely to possess a BFAH motif. The user

may vary the periodicity from that of an alpha-helix (100 degrees ), to those of

other secondary structures such as beta sheets and 3(10) helices. The program can

also plot the periodicity of basic residues in a protein sequence using a Fourier

transformation. The procedure has been used to characterize the presence of BFAHs

in the N-terminal extensions of the eukaryotic aminoacyl-tRNA synthetases and to

indicate the presence of a BFAH in the tRNA binding site of alanyl-tRNA

synthetase.AVAILABILITY: www.scanmoment.org.

PMCID: PMC2655046

PMID: 19293994

2306. Bioinformation. 2009;3(7):287-8. Epub 2009 Feb 26.

Finding Alu in primate genomes with AF-1.

Shankar R(1), Kataria B, Mukerji M.

Author information:

(1)Department of Bioinformatics & Structural Biology, School of Life Sciences &

Biotechnology, Indian Institute of Advanced Research, Gandhinagar, Gujarat,

India. ravish@iiar.res.in

Repetitive sequences occupy more than 40% of the human genome which is much

larger compared to the 2% occupied by the coding DNA. Amongst these Alu elements

are the second largest class of repeats, occupying nearly 10% of the whole

genome. Alus have been implicated in many genomic processes, sometimes giving

rise to aberrations while many times playing as silent player in genomic and

regulatory evolution. Here we present a web server, AF1, exclusively developed

for finding Alu like elements. Besides alignment based methodology, this server

utilizes probabilistic scanning to find more diverged elements and employs a more

precise way of element classification based on unequal weighting of sequence

through sequence encoding.AVAILABILITY: AF1 is freely available at

http://software.iiar.res.in/af1/. The standalone is also available for download.

PMCID: PMC2652563

PMID: 19293991

2307. Kokuritsu Iyakuhin Shokuhin Eisei Kenkyusho Hokoku. 2009;(127):44-9.

[Major revision of the allergen database for food safety (ADFS) and validation of

the motif-based allergenicity prediction tool].

[Article in Japanese]

Nakamura R(1), Nakamura R, Teshima R.

Author information:

(1)ryosnak@nihs.go.jp

We have been maintaining an integral web server system, the Allergen Database for

Food Safety (ADFS), since 2005 (http://allergen.nihs.go.jp/ADFS/). Recently, a

group at the University of Nebraska-Lincoln released a new version of an allergen

database, AllergenOnline. This database includes more than 1,300 allergens, all

of which have been peer-reviewed by an international board of allergology

experts. Here, we have totally revised the dataset of the ADFS by comparing it

with that of AllergenOnline to improve the reliability of our allergen data.

Moreover, the performance of our web-based tool for predicting new allergens

(motif-based method), which was developed according to a theory proposed by

Stadler & Stadler (2003), was validated using three methods. As a result of the

integration of this allergen data, the number of (iso)allergens in the ADFS has

increased to 1340, and epitope information is now available for 76 allergens.

Using model datasets, the precision, recall, and specificity of our motif-based

allergenicity prediction tool was proved to be 100.0%, 99.4%, and 100.0%,

respectively. These results were similar to those for the original motif-based

prediction model that was previously reported and are much better than those of

the method recommended by FAO/WHO, especially with regard to the precision of

predictions.

PMID: 20306706 [Indexed for MEDLINE]

2308. Nat Protoc. 2009;4(7):1073-81. doi: 10.1038/nprot.2009.86. Epub 2009 Jun 25.

Predicting the effects of coding non-synonymous variants on protein function

using the SIFT algorithm.

Kumar P(1), Henikoff S, Ng PC.

Author information:

(1)Department of Genomic Medicine, J. Craig Venter Institute, San Diego,

California, USA.

The effect of genetic mutation on phenotype is of significant interest in

genetics. The type of genetic mutation that causes a single amino acid

substitution (AAS) in a protein sequence is called a non-synonymous single

nucleotide polymorphism (nsSNP). An nsSNP could potentially affect the function

of the protein, subsequently altering the carrier's phenotype. This protocol

describes the use of the 'Sorting Tolerant From Intolerant' (SIFT) algorithm in

predicting whether an AAS affects protein function. To assess the effect of a

substitution, SIFT assumes that important positions in a protein sequence have

been conserved throughout evolution and therefore substitutions at these

positions may affect protein function. Thus, by using sequence homology, SIFT

predicts the effects of all possible substitutions at each position in the

protein sequence. The protocol typically takes 5-20 min, depending on the input.

SIFT is available as an online tool (http://sift.jcvi.org).

DOI: 10.1038/nprot.2009.86

PMID: 19561590 [Indexed for MEDLINE]

2309. Nucleic Acids Res. 2009 Jan;37(Database issue):D1025-8. doi: 10.1093/nar/gkn966.

The YH database: the first Asian diploid genome database.

Li G(1), Ma L, Song C, Yang Z, Wang X, Huang H, Li Y, Li R, Zhang X, Yang H, Wang

J, Wang J.

Author information:

(1)Beijing Genomics Institute at Shenzhen 518083, China.

The YH database is a server that allows the user to easily browse and download

data from the first Asian diploid genome. The aim of this platform is to

facilitate the study of this Asian genome and to enable improved organization and

presentation large-scale personal genome data. Powered by GBrowse, we illustrate

here the genome sequences, SNPs, and sequencing reads in the MapView. The

relationships between phenotype and genotype can be searched by location, dbSNP

ID, HGMD ID, gene symbol and disease name. A BLAST web service is also provided

for the purpose of aligning query sequence against YH genome consensus. The YH

database is currently one of the three personal genome database, organizing the

original data and analysis results in a user-friendly interface, which is an

endeavor to achieve fundamental goals for establishing personal medicine. The

database is available at http://yh.genomics.org.cn.

DOI: 10.1093/nar/gkn966

PMCID: PMC2686535

PMID: 19073702 [Indexed for MEDLINE]

2310. Nucleic Acids Res. 2009 Jan;37(Database issue):D598-602. doi: 10.1093/nar/gkn864.

Epub 2008 Nov 5.

Bionemo: molecular information on biodegradation metabolism.

Carbajosa G(1), Trigo A, Valencia A, Cases I.

Author information:

(1)Structural Biology and Biocomputing Programme, Spanish National Cancer

Research Centre (CNIO), Melchor Fernández Almagro, 3, E-28029, Madrid, Spain.

Bionemo (http://bionemo.bioinfo.cnio.es) stores manually curated information

about proteins and genes directly implicated in the Biodegradation metabolism.

When possible, the database includes information on sequence, domains and

structures for proteins; and sequence, regulatory elements and transcription

units for genes. Thus, Bionemo is a unique resource that complements other

biodegradation databases such as the University of Minessota

Biocatalysis/Biodegradation Database, or Metarouter, which focus more on the

biochemical aspects of biodegradation than in the nature of the biomolecules

carrying out the reactions. Bionemo has been built by manually associating

sequences database entries to biodegradation reactions, using the information

extracted from published articles. Information on transcription units and their

regulation was also extracted from the literature for biodegradation genes, and

linked to the underlying biochemical network. In its current version, Bionemo

contains sequence information for 324 reactions and transcription regulation

information for more than 100 promoters and 100 transcription factors. The

information in the Bionemo database is available via a web server and the full

database is also downloadable as a PostgresSQL dump. To facilitate the

programmatic use of the information contained in the database, an object-oriented

Perl API is also provided.

DOI: 10.1093/nar/gkn864

PMCID: PMC2686592

PMID: 18986994 [Indexed for MEDLINE]

2311. Nucleic Acids Res. 2009 Jan;37(Database issue):D229-32. doi: 10.1093/nar/gkn808.

Epub 2008 Oct 31.

SMART 6: recent updates and new developments.

Letunic I(1), Doerks T, Bork P.

Author information:

(1)EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany.

Simple modular architecture research tool (SMART) is an online tool

(http://smart.embl.de/) for the identification and annotation of protein domains.

It provides a user-friendly platform for the exploration and comparative study of

domain architectures in both proteins and genes. The current release of SMART

contains manually curated models for 784 protein domains. Recent developments

were focused on further data integration and improving user friendliness. The

underlying protein database based on completely sequenced genomes was greatly

expanded and now includes 630 species, compared to 191 in the previous release.

As an initial step towards integrating information on biological pathways into

SMART, our domain annotations were extended with data on metabolic pathways and

links to several pathways resources. The interaction network view was completely

redesigned and is now available for more than 2 million proteins. In addition to

the standard web access to the database, users can now query SMART using

distributed annotation system (DAS) or through a simple object access protocol

(SOAP) based web service.

DOI: 10.1093/nar/gkn808

PMCID: PMC2686533

PMID: 18978020 [Indexed for MEDLINE]

2312. Nucleic Acids Res. 2009 Jan;37(Database issue):D323-7. doi: 10.1093/nar/gkn822.

Epub 2008 Oct 29.

The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein

structures.

Goldenberg O(1), Erez E, Nimrod G, Ben-Tal N.

Author information:

(1)Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv

University, Israel.

ConSurf-DB is a repository for evolutionary conservation analysis of the proteins

of known structures in the Protein Data Bank (PDB). Sequence homologues of each

of the PDB entries were collected and aligned using standard methods. The

evolutionary conservation of each amino acid position in the alignment was

calculated using the Rate4Site algorithm, implemented in the ConSurf web server.

The algorithm takes into account the phylogenetic relations between the aligned

proteins and the stochastic nature of the evolutionary process explicitly.

Rate4Site assigns a conservation level for each position in the multiple sequence

alignment using an empirical Bayesian inference. Visual inspection of the

conservation patterns on the 3D structure often enables the identification of key

residues that comprise the functionally important regions of the protein. The

repository is updated with the latest PDB entries on a monthly basis and will be

rebuilt annually. ConSurf-DB is available online at http://consurfdb.tau.ac.il/

DOI: 10.1093/nar/gkn822

PMCID: PMC2686473

PMID: 18971256 [Indexed for MEDLINE]

2313. Nucleic Acids Res. 2009 Jan;37(Database issue):D393-5. doi: 10.1093/nar/gkn769.

Epub 2008 Oct 23.

Voronoia: analyzing packing in protein structures.

Rother K(1), Hildebrand PW, Goede A, Gruening B, Preissner R.

Author information:

(1)International Institute for Molecular and Cell Biology, ul. ks. Trojdena 4,

02-109 Warszawa, Poland. krother@genesilico.pl

The packing of protein atoms is an indicator for their stability and

functionality, and applied in determining thermostability, in protein design,

ligand binding and to identify flexible regions in proteins. Here, we present

Voronoia, a database of atomic-scale packing data for protein 3D structures. It

is based on an improved Voronoi Cell algorithm using hyperboloid interfaces to

construct atomic volumes, and to resolve solvent-accessible and -inaccessible

regions of atoms. The database contains atomic volumes, local packing densities

and interior cavities calculated for 61 318 biological units from the PDB. A

report for each structure summarizes the packing by residue and atom types, and

lists the environment of interior cavities. The packing data are compared to a

nonredundant set of structures from SCOP superfamilies. Both packing densities

and cavities can be visualized in the 3D structures by the Jmol plugin.

Additionally, PDB files can be submitted to the Voronoia server for calculation.

This service performs calculations for most full-atomic protein structures within

a few minutes. For batch jobs, a standalone version of the program with an

optional PyMOL plugin is available for download. The database can be freely

accessed at: http://bioinformatics.charite.de/voronoia.

DOI: 10.1093/nar/gkn769

PMCID: PMC2686436

PMID: 18948293 [Indexed for MEDLINE]

2314. Nucleic Acids Res. 2009 Jan;37(Database issue):D347-54. doi: 10.1093/nar/gkn791.

Epub 2008 Oct 23.

MODBASE, a database of annotated comparative protein structure models and

associated resources.

Pieper U(1), Eswar N, Webb BM, Eramian D, Kelly L, Barkan DT, Carter H, Mankoo P,

Karchin R, Marti-Renom MA, Davis FP, Sali A.

Author information:

(1)Department of Bioengineering and Therapeutic Sciences, Department of

Pharmaceutical Chemistry, University of California at San Francisco, 1700 4th

Street, San Francisco, CA 94158, USA.

MODBASE (http://salilab.org/modbase) is a database of annotated comparative

protein structure models. The models are calculated by MODPIPE, an automated

modeling pipeline that relies primarily on MODELLER for fold assignment,

sequence-structure alignment, model building and model assessment

(http:/salilab.org/modeller). MODBASE currently contains 5,152,695 reliable

models for domains in 1,593,209 unique protein sequences; only models based on

statistically significant alignments and/or models assessed to have the correct

fold are included. MODBASE also allows users to calculate comparative models on

demand, through an interface to the MODWEB modeling server

(http://salilab.org/modweb). Other resources integrated with MODBASE include

databases of multiple protein structure alignments (DBAli), structurally defined

ligand binding sites (LIGBASE), predicted ligand binding sites (AnnoLyze),

structurally defined binary domain interfaces (PIBASE) and annotated single

nucleotide polymorphisms and somatic mutations found in human proteins (LS-SNP,

LS-Mut). MODBASE models are also available through the Protein Model Portal

(http://www.proteinmodelportal.org/).

DOI: 10.1093/nar/gkn791

PMCID: PMC2686492

PMID: 18948282 [Indexed for MEDLINE]

2315. Nucleic Acids Res. 2009 Jan;37(Database issue):D731-7. doi: 10.1093/nar/gkn645.

Epub 2008 Oct 2.

PhyloPat: an updated version of the phylogenetic pattern database contains gene

neighborhood.

Hulsen T(1), Groenen PM, de Vlieg J, Alkema W.

Author information:

(1)Computational Drug Discovery, CMBI, NCMLS, Radboud University Nijmegen Medical

Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands. t.hulsen@cmbi.ru.nl

Phylogenetic patterns show the presence or absence of certain genes in a set of

full genomes derived from different species. They can also be used to determine

sets of genes that occur only in certain evolutionary branches. Previously, we

presented a database named PhyloPat which allows the complete Ensembl gene

database to be queried using phylogenetic patterns. Here, we describe an updated

version of PhyloPat which can be queried by an improved web server. We used a

single linkage clustering algorithm to create 241,697 phylogenetic lineages,

using all the orthologies provided by Ensembl v49. PhyloPat offers the

possibility of querying with binary phylogenetic patterns or regular expressions,

or through a phylogenetic tree of the 39 included species. Users can also input a

list of Ensembl, EMBL, EntrezGene or HGNC IDs to check which phylogenetic lineage

any gene belongs to. A link to the FatiGO web interface has been incorporated in

the HTML output. For each gene, the surrounding genes on the chromosome, color

coded according to their phylogenetic lineage can be viewed, as well as FASTA

files of the peptide sequences of each lineage. Furthermore, lists of

omnipresent, polypresent, oligopresent and anticorrelating genes have been

included. PhyloPat is freely available at http://www.cmbi.ru.nl/phylopat.

DOI: 10.1093/nar/gkn645

PMCID: PMC2686476

PMID: 18832367 [Indexed for MEDLINE]

2316. Nucleic Acids Res. 2009 Jan;37(Database issue):D686-9. doi: 10.1093/nar/gkn648.

Epub 2008 Oct 2.

CleanEST: a database of cleansed EST libraries.

Lee B(1), Shin G.

Author information:

(1)Korean BioInformation Center, KRIBB, Daejeon 305-817, Korea. bulee@kribb.re.kr

The EST division of GenBank, dbEST, is widely used in many applications such as

gene discovery and verification of exon-intron structure. However, the use of EST

sequences in the dbEST libraries is often hampered by inconsistent terminology

used to describe the library sources and by the presence of contaminated

sequences. Here, we describe CleanEST, a novel database server that classified

dbEST libraries and removes contaminants. We classified all dbEST libraries

according to species and sequencing center. In addition, we further classified

human EST libraries by anatomical and pathological systems according to eVOC

ontologies. For each dbEST library, we provide two different cleansed sequences:

'pre-cleansed' and 'user-cleansed'. To generate pre-cleansed sequences, we

cleansed sequences in dbEST by alignment of EST sequences against well-known

contamination sources: UniVec, Escherichia coli, mitochondria and chloroplast

(for plant). To provide user-cleansed sequences, we built an automatic

user-cleansing pipeline, in which sequences of a user-selected library are

cleansed on-the-fly according to user-selected options. The server is available

at http://cleanest.kobic.re.kr/ and the database is updated monthly.

DOI: 10.1093/nar/gkn648

PMCID: PMC2686460

PMID: 18832365 [Indexed for MEDLINE]

2317. Protein Pept Lett. 2009;16(12):1478-84.

Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular

localization of Gram-positive bacterial proteins.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong

University, 800 Dongchuan Road, Shanghai, China.

In this paper, a new predictor called "Gpos-mPLoc", is developed for identifying

the subcellular localization of Gram positive bacterial proteins by fusing the

information of gene ontology, as well as the functional domain information and

sequential evolution information. Compared with the old Gpos-PLoc, the new

predictor is much more powerful and flexible. Particularly, it also has the

capacity to deal with multiple-location proteins as indicated by the character

"m" in front of "PLoc" of its name. For a newly-constructed stringent benchmark

dataset in which none of included proteins has > 25% pairwise sequence identity

to any other in a same subset (location), the overall jackknife success rate

achieved by Gpos-mPLoc was 82.2%, which was about 10% higher than the

corresponding rate by the Gpos-PLoc. As a user friendly web-server, Gpos-mPLoc is

freely accessible at http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi/.

PMID: 20001911 [Indexed for MEDLINE]

2318. Protein Pept Lett. 2009;16(12):1447-54.

Robust prediction of B-factor profile from sequence using two-stage SVR based on

random forest feature selection.

Pan XY(1), Shen HB.

Author information:

(1)Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong

University, 800 Dongchuan Road, Shanghai, China.

B-factor is highly correlated with protein internal motion, which is used to

measure the uncertainty in the position of an atom within a crystal structure.

Although the rapid progress of structural biology in recent years makes more

accurate protein structures available than ever, with the avalanche of new

protein sequences emerging during the post-genomic Era, the gap between the known

protein sequences and the known protein structures becomes wider and wider. It is

urgent to develop automated methods to predict B-factor profile from the amino

acid sequences directly, so as to be able to timely utilize them for basic

research. In this article, we propose a novel approach, called PredBF, to predict

the real value of B-factor. We firstly extract both global and local features

from the protein sequences as well as their evolution information, then the

random forests feature selection is applied to rank their importance and the most

important features are inputted to a two-stage support vector regression (SVR)

for prediction, where the initial predicted outputs from the 1(st) SVR are

further inputted to the 2nd layer SVR for final refinement. Our results have

revealed that a systematic analysis of the importance of different features makes

us have deep insights into the different contributions of features and is very

necessary for developing effective B-factor prediction tools. The two-layer SVR

prediction model designed in this study further enhanced the robustness of

predicting the B-factor profile. As a web server, PredBF is freely available at:

http://www.csbio.sjtu.edu.cn/bioinf/PredBF for academic use.

PMID: 20001907 [Indexed for MEDLINE]

2319. Protein Pept Lett. 2009;16(8):977-83.

Improved prediction of lysine acetylation by support vector machines.

Li S(1), Li H, Li M, Shyr Y, Xie L, Li Y.

Author information:

(1)Bio-X Life Science Research Center and School of Life Science and

Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.

Reversible acetylation on lysine residues, a crucial post-translational

modification (PTM) for both histone and non-histone proteins, governs many

central cellular processes. Due to limited data and lack of a clear acetylation

consensus sequence, little research has focused on prediction of lysine

acetylation sites. Incorporating almost all currently available lysine

acetylation information, and using the support vector machine (SVM) method along

with coding schema for protein sequence coupling patterns, we propose here a

novel lysine acetylation prediction algorithm: LysAcet. When compared with other

methods or existing tools, LysAcet is the best predictor of lysine acetylation,

with K-fold (5- and 10-) and jackknife cross-validation accuracies of 75.89%,

76.73%, and 77.16%, respectively. LysAcet's superior predictive accuracy is

attributed primarily to the use of sequence coupling patterns, which describe the

relative position of two amino acids. LysAcet contributes to the limited PTM

prediction research on lysine epsilon-acetylation, and may serve as a

complementary in-silicon approach for exploring acetylation on proteomes. An

online web server is freely available at http://www.biosino.org/LysAcet/.

PMID: 19689425 [Indexed for MEDLINE]

2320. Proteins. 2009;77 Suppl 9:185-90. doi: 10.1002/prot.22491.

Prediction of global and local model quality in CASP8 using the ModFOLD server.

McGuffin LJ(1).

Author information:

(1)School of Biological Sciences, University of Reading Whiteknights, Reading RG6

6AS, United Kingdom. l.j.mcguffin@reading.ac.uk

The development of effective methods for predicting the quality of

three-dimensional (3D) models is fundamentally important for the success of

tertiary structure (TS) prediction strategies. Since CASP7, the Quality

Assessment (QA) category has existed to gauge the ability of various model

quality assessment programs (MQAPs) at predicting the relative quality of

individual 3D models. For the CASP8 experiment, automated predictions were

submitted in the QA category using two methods from the ModFOLD server-ModFOLD

version 1.1 and ModFOLDclust. ModFOLD version 1.1 is a single-model machine

learning based method, which was used for automated predictions of global model

quality (QMODE1). ModFOLDclust is a simple clustering based method, which was

used for automated predictions of both global and local quality (QMODE2). In

addition, manual predictions of model quality were made using ModFOLD version

2.0--an experimental method that combines the scores from ModFOLDclust and

ModFOLD v1.1. Predictions from the ModFOLDclust method were the most successful

of the three in terms of the global model quality, whilst the ModFOLD v1.1 method

was comparable in performance to other single-model based methods. In addition,

the ModFOLDclust method performed well at predicting the per-residue, or local,

model quality scores. Predictions of the per-residue errors in our own 3D models,

selected using the ModFOLD v2.0 method, were also the most accurate compared with

those from other methods. All of the MQAPs described are publicly accessible via

the ModFOLD server at: http://www.reading.ac.uk/bioinf/ModFOLD/. The methods are

also freely available to download from:

http://www.reading.ac.uk/bioinf/downloads/.

Copyright 2009 Wiley-Liss, Inc.

DOI: 10.1002/prot.22491

PMID: 19585661 [Indexed for MEDLINE]

2321. Stud Health Technol Inform. 2009;143:93-8.

Technical and architectural issues in deploying electronic health records (EHRs)

over the WWW.

Armstrong B(1), Kushniruk A, Joe R, Borycki E.

Author information:

(1)University of Victoria, Victoria, BC, Canada.

In this paper technical and architectural issues are described in deploying

electronic health records (EHRs) over the WWW. The project described involved

deployment of EHRs that have been designed to serve in the education of health

professionals and health/biomedical informaticians. In order to allow for

ubiquitous access to a range of EHRs remotely an architecture was designed with

three layers: (a) the "Internet" or remote user access layer (2) the "Perimeter

Network", or middle firewall security and authentication layer (3) the "HINF EHR

Network", consisting of the internal servers hosting EHR applications and

databases. The approaches allow for a large number of remote users running a

range of operating systems to access the educational EHRs from any location

remotely. Virtual machine (VM) technology is employed to allow multiple versions

and platforms of operating systems to be installed side-by-side on a single

server. Security, technical and budgetary considerations are described as well as

past and current applications of the architecture for a number of projects for

the education of health professionals in the area of electronic health records.

PMID: 19380921 [Indexed for MEDLINE]

2322. Bioinformatics. 2008 Dec 15;24(24):2930-1. doi: 10.1093/bioinformatics/btn559.

Epub 2008 Nov 18.

Motif Tool Manager: a web-based framework for motif discovery.

Phan V(1), Furlotte NA.

Author information:

(1)Department of Computer Science, The Bioinformatics Program and The W. Harry

Feinstone Center for Genomics Research, The University of Memphis, Memphis, TN

38152, USA. vphan@memphis.edu

MOTIVATION: Motif Tool Manager is a web-based framework for comparing and

combining different approaches to discover novel DNA motifs. It comes with a set

of five well-known approaches to motif discovery. It provides an easy mechanism

for adding new motif finding tools to the framework through a web-interface and a

minimal setup of the tools on the server. Users can execute the tools through the

web-based framework and compare results from such executions. The framework

provides a basic mechanism for identifying the most similar motif candidates

found by a majority of themotif finding tools.

AVAILABILITY: http://cetus.cs.memphis.edu/motif

DOI: 10.1093/bioinformatics/btn559

PMID: 19017656 [Indexed for MEDLINE]

2323. Bioinformatics. 2008 Dec 15;24(24):2938-9. doi: 10.1093/bioinformatics/btn564.

Epub 2008 Oct 30.

SNAP: a web-based tool for identification and annotation of proxy SNPs using

HapMap.

Johnson AD(1), Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI.

Author information:

(1)The Framingham Heart Study of the National Heart, Lung, and Blood Institute of

the National Institutes of Health, USA.

SUMMARY: The interpretation of genome-wide association results is confounded by

linkage disequilibrium between nearby alleles. We have developed a flexible

bioinformatics query tool for single-nucleotide polymorphisms (SNPs) to identify

and to annotate nearby SNPs in linkage disequilibrium (proxies) based on HapMap.

By offering functionality to generate graphical plots for these data, the SNAP

server will facilitate interpretation and comparison of genome-wide association

study results, and the design of fine-mapping experiments (by delineating genomic

regions harboring associated variants and their proxies).

AVAILABILITY: SNAP server is available at http://www.broad.mit.edu/mpg/snap/.

DOI: 10.1093/bioinformatics/btn564

PMCID: PMC2720775

PMID: 18974171 [Indexed for MEDLINE]

2324. Bioinformatics. 2008 Dec 15;24(24):2923-5. doi: 10.1093/bioinformatics/btn541.

Epub 2008 Oct 30.

iMapper: a web application for the automated analysis and mapping of insertional

mutagenesis sequence data against Ensembl genomes.

Kong J(1), Zhu F, Stalker J, Adams DJ.

Author information:

(1)Experimental Cancer Genetics, The Wellcome Trust Sanger Institute, Hinxton,

Cambs CB101HH, UK.

SUMMARY: Insertional mutagenesis is a powerful method for gene discovery. To

identify the location of insertion sites in the genome linker based polymerase

chain reaction (PCR) methods (such as splinkerette-PCR) may be employed. We have

developed a web application called iMapper (Insertional Mutagenesis Mapping and

Analysis Tool) for the efficient analysis of insertion site sequence reads

against vertebrate and invertebrate Ensembl genomes. Taking linker based

sequences as input, iMapper scans and trims the sequence to remove the linker and

sequences derived from the insertional mutagen. The software then identifies and

removes contaminating sequences derived from chimeric genomic fragments, vector

or the transposon concatamer and then presents the clipped sequence reads to a

sequence mapping server which aligns them to an Ensembl genome. Insertion sites

can then be navigated in Ensembl in the context of genomic features such as gene

structures. iMapper also generates test-based format for nucleic acid or protein

sequences (FASTA) and generic file format (GFF) files of the clipped sequence

reads and provides a graphical overview of the mapped insertion sites against a

karyotype. iMapper is designed for high-throughput applications and can

efficiently process thousands of DNA sequence reads.

AVAILABILITY: iMapper is web based and can be accessed at

http://www.sanger.ac.uk/cgi-bin/teams/team113/imapper.cgi.

DOI: 10.1093/bioinformatics/btn541

PMCID: PMC2639305

PMID: 18974167 [Indexed for MEDLINE]

2325. Bioinformatics. 2008 Dec 15;24(24):2928-9. doi: 10.1093/bioinformatics/btn550.

Epub 2008 Oct 22.

SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology.

Viklund H(1), Bernsel A, Skwark M, Elofsson A.

Author information:

(1)Department of Biochemistry and Biophysics, Center for Biomembrane Research,

Stockholm Bioinformatics Center, The Arrhenius Laboratories for Natural Sciences,

Stockholm University, SE-10691 Stockholm, Sweden.

SUMMARY: SPOCTOPUS is a method for combined prediction of signal peptides and

membrane protein topology, suitable for genome-scale studies. Its objective is to

minimize false predictions of transmembrane regions as signal peptides and vice

versa. We provide a description of the SPOCTOPUS algorithm together with a

performance evaluation where SPOCTOPUS compares favorably with state-of-the-art

methods for signal peptide and topology predictions.

AVAILABILITY: SPOCTOPUS is available as a web server and both the source code and

benchmark data are available for download at http://octopus.cbr.su.se/

DOI: 10.1093/bioinformatics/btn550

PMID: 18945683 [Indexed for MEDLINE]

2326. Structure. 2008 Dec 10;16(12):1755-63. doi: 10.1016/j.str.2008.10.017.

Detection of functionally important regions in "hypothetical proteins" of known

structure.

Nimrod G(1), Schushan M, Steinberg DM, Ben-Tal N.

Author information:

(1)Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv

University, 69978 Tel Aviv, Israel.

Structural genomics initiatives provide ample structures of "hypothetical

proteins" (i.e., proteins of unknown function) at an ever increasing rate.

However, without function annotation, this structural goldmine is of little use

to biologists who are interested in particular molecular systems. To this end, we

used (an improved version of) the PatchFinder algorithm for the detection of

functional regions on the protein surface, which could mediate its interactions

with, e.g., substrates, ligands, and other proteins. Examination, using a data

set of annotated proteins, showed that PatchFinder outperforms similar methods.

We collected 757 structures of hypothetical proteins and their predicted

functional regions in the N-Func database. Inspection of several of these regions

demonstrated that they are useful for function prediction. For example, we

suggested an interprotein interface and a putative nucleotide-binding site. A

web-server implementation of PatchFinder and the N-Func database are available at

http://patchfinder.tau.ac.il/.

DOI: 10.1016/j.str.2008.10.017

PMID: 19081051 [Indexed for MEDLINE]

2327. Retrovirology. 2008 Dec 4;5:110. doi: 10.1186/1742-4690-5-110.

HIV-1 coreceptor usage prediction without multiple alignments: an application of

string kernels.

Boisvert S(1), Marchand M, Laviolette F, Corbeil J.

Author information:

(1)Centre de recherche du centre hospitalier de l'Université Laval, Québec,

Canada. Sebastien.Boisvert.3@ulaval.ca

BACKGROUND: Human immunodeficiency virus type 1 (HIV-1) infects cells by means of

ligand-receptor interactions. This lentivirus uses the CD4 receptor in

conjunction with a chemokine coreceptor, either CXCR4 or CCR5, to enter a target

cell. HIV-1 is characterized by high sequence variability. Nonetheless, within

this extensive variability, certain features must be conserved to define

functions and phenotypes. The determination of coreceptor usage of HIV-1, from

its protein envelope sequence, falls into a well-studied machine learning problem

known as classification. The support vector machine (SVM), with string kernels,

has proven to be very efficient for dealing with a wide class of classification

problems ranging from text categorization to protein homology detection. In this

paper, we investigate how the SVM can predict HIV-1 coreceptor usage when it is

equipped with an appropriate string kernel.

RESULTS: Three string kernels were compared. Accuracies of 96.35% (CCR5) 94.80%

(CXCR4) and 95.15% (CCR5 and CXCR4) were achieved with the SVM equipped with the

distant segments kernel on a test set of 1425 examples with a classifier built on

a training set of 1425 examples. Our datasets are built with Los Alamos National

Laboratory HIV Databases sequences. A web server is available at

http://genome.ulaval.ca/hiv-dskernel.

CONCLUSION: We examined string kernels that have been used successfully for

protein homology detection and propose a new one that we call the distant

segments kernel. We also show how to extract the most relevant features for HIV-1

coreceptor usage. The SVM with the distant segments kernel is currently the best

method described.

DOI: 10.1186/1742-4690-5-110

PMCID: PMC2637298

PMID: 19055831 [Indexed for MEDLINE]

2328. BMC Struct Biol. 2008 Dec 3;8:52. doi: 10.1186/1472-6807-8-52.

SELECTpro: effective protein model selection using a structure-based energy

function resistant to BLUNDERs.

Randall A(1), Baldi P.

Author information:

(1)School of Information and Computer Sciences, University of California, Irvine,

CA 92697, USA. arandall@ics.uci.edu

BACKGROUND: Protein tertiary structure prediction is a fundamental problem in

computational biology and identifying the most native-like model from a set of

predicted models is a key sub-problem. Consensus methods work well when the

redundant models in the set are the most native-like, but fail when the most

native-like model is unique. In contrast, structure-based methods score models

independently and can be applied to model sets of any size and redundancy level.

Additionally, structure-based methods have a variety of important applications

including analogous fold recognition, refinement of sequence-structure

alignments, and de novo prediction. The purpose of this work was to develop a

structure-based model selection method based on predicted structural features

that could be applied successfully to any set of models.

RESULTS: Here we introduce SELECTpro, a novel structure-based model selection

method derived from an energy function comprising physical, statistical, and

predicted structural terms. Novel and unique energy terms include predicted

secondary structure, predicted solvent accessibility, predicted contact map,

beta-strand pairing, and side-chain hydrogen bonding.SELECTpro participated in

the new model quality assessment (QA) category in CASP7, submitting predictions

for all 95 targets and achieved top results. The average difference in GDT-TS

between models ranked first by SELECTpro and the most native-like model was 5.07.

This GDT-TS difference was less than 1% of the GDT-TS of the most native-like

model for 18 targets, and less than 10% for 66 targets. SELECTpro also ranked the

single most native-like first for 15 targets, in the top five for 39 targets, and

in the top ten for 53 targets, more often than any other method. Because the

ranking metric is skewed by model redundancy and ignores poor models with a

better ranking than the most native-like model, the BLUNDER metric is introduced

to overcome these limitations. SELECTpro is also evaluated on a recent benchmark

set of 16 small proteins with large decoy sets of 12500 to 20000 models for each

protein, where it outperforms the benchmarked method (I-TASSER).

CONCLUSION: SELECTpro is an effective model selection method that scores models

independently and is appropriate for use on any model set. SELECTpro is available

for download as a stand alone application at:

http://www.igb.uci.edu/~baldig/selectpro.html. SELECTpro is also available as a

public server at the same site.

DOI: 10.1186/1472-6807-8-52

PMCID: PMC2667183

PMID: 19055744 [Indexed for MEDLINE]

2329. Bioinformatics. 2008 Dec 1;24(23):2698-705. doi: 10.1093/bioinformatics/btn518.

Epub 2008 Oct 7.

A novel method for comparing topological models of protein structures enhanced

with ligand information.

Veeramalai M(1), Gilbert D.

Author information:

(1)Bioinformatics Research Centre, Department of Computing Science, University of

Glasgow, Glasgow G12 8QQ, UK. mallikav@burnham.org

We introduce TOPS+ strings, a highly abstract string-based model of protein

topology that permits efficient computation of structure comparison, and can

optionally represent ligand information. In this model, we consider loops as

secondary structure elements (SSEs) as well as helices and strands; in addition

we represent ligands as first class objects. Interactions between SSEs and

between SSEs and ligands are described by incoming/outgoing arcs and ligand arcs,

respectively; and SSEs are annotated with arc interaction direction and type. We

are able to abstract away from the ligands themselves, to give a model

characterized by a regular grammar rather than the context sensitive grammar of

the original TOPS model. Our TOPS+ strings model is sufficiently descriptive to

obtain biologically meaningful results and has the advantage of permitting fast

string-based structure matching and comparison as well as avoiding issues of

Non-deterministic Polynomial time (NP)-completeness associated with graph

problems. Our structure comparison method is computationally more efficient in

identifying distantly related proteins than BLAST, CLUSTALW, SSAP and TOPS

because of the compact and abstract string-based representation of protein

structure which records both topological and biochemical information including

the functionally important loop regions of the protein structures. The accuracy

of our comparison method is comparable with that of TOPS. Also, we have

demonstrated that our TOPS+ strings method out-performs the TOPS method for the

ligand-dependent protein structures and provides biologically meaningful

results.AVAILABILITY: The TOPS+ strings comparison server is available from

http://balabio.dcs.gla.ac.uk/mallika/WebTOPS/topsplus.html.

DOI: 10.1093/bioinformatics/btn518

PMID: 18842602 [Indexed for MEDLINE]

2330. Bioinformatics. 2008 Dec 1;24(23):2786-7. doi: 10.1093/bioinformatics/btn522.

Epub 2008 Oct 7.

A flexible forward simulator for populations subject to selection and demography.

Hernandez RD(1).

Author information:

(1)Biological Statistics and Computational Biology, Cornell University, Ithaca,

NY 14850, USA. rhernandez@uchicago.edu

This article introduces a new forward population genetic simulation program that

can efficiently generate samples from populations with complex demographic

histories under various models of natural selection. The program (SFS\_CODE) is

highly flexible, allowing the user to simulate realistic genomic regions with

several loci evolving according to a variety of mutation models (from simple to

context-dependent), and allows for insertions and deletions. Each locus can be

annotated as either coding or non-coding, sex-linked or autosomal, selected or

neutral, and have an arbitrary linkage structure (from completely linked to

independent).AVAILABILITY: The source code (written in the C programming

language) is available at http://sfscode.sourceforge.net, and a web server

(http://cbsuapps.tc.cornell.edu/sfscode.aspx) allows the user to perform

simulations using the high-performance computing cluster hosted by the Cornell

University Computational Biology Service Unit.

DOI: 10.1093/bioinformatics/btn522

PMCID: PMC2639268

PMID: 18842601 [Indexed for MEDLINE]

2331. Bioinformatics. 2008 Dec 1;24(23):2780-1. doi: 10.1093/bioinformatics/btn507.

Epub 2008 Sep 25.

Searching protein structure databases with DaliLite v.3.

Holm L(1), Kääriäinen S, Rosenström P, Schenkel A.

Author information:

(1)Department of Biological and Environmental Sciences, and Institute of

Biotechnology, P.O.Box 56 (Viikinkaari 5), 00014 University of Helsinki, Finland.

liisa.holm@helsinki.fi

The Red Queen said, 'It takes all the running you can do, to keep in the same

place.' Lewis CarrolMOTIVATION: Newly solved protein structures are routinely

scanned against structures already in the Protein Data Bank (PDB) using Internet

servers. In favourable cases, comparing 3D structures may reveal biologically

interesting similarities that are not detectable by comparing sequences. The

number of known structures continues to grow exponentially. Sensitive-thorough

but slow-search algorithms are challenged to deliver results in a reasonable

time, as there are now more structures in the PDB than seconds in a day. The

brute-force solution would be to distribute the individual comparisons on a

massively parallel computer. A frugal solution, as implemented in the Dali

server, is to reduce the total computational cost by pruning search space using

prior knowledge about the distribution of structures in fold space. This note

reports paradigm revisions that enable maintaining such a knowledge base

up-to-date on a PC.

AVAILABILITY: The Dali server for protein structure database searching at

http://ekhidna.biocenter.helsinki.fi/dali\_server is running DaliLite v.3. The

software can be downloaded for academic use from

http://ekhidna.biocenter.helsinki.fi/dali\_lite/downloads/v3.

DOI: 10.1093/bioinformatics/btn507

PMCID: PMC2639270

PMID: 18818215 [Indexed for MEDLINE]

2332. Bioinformatics. 2008 Dec 1;24(23):2726-32. doi: 10.1093/bioinformatics/btn452.

Epub 2008 Sep 2.

MeltDB: a software platform for the analysis and integration of metabolomics

experiment data.

Neuweger H(1), Albaum SP, Dondrup M, Persicke M, Watt T, Niehaus K, Stoye J,

Goesmann A.

Author information:

(1)International NRW Graduate School in Bioinformatics and Genome Research,

Bielefeld University, Germany. hneuwege@cebitec.uni-bielefeld.de

MOTIVATION: The recent advances in metabolomics have created the potential to

measure the levels of hundreds of metabolites which are the end products of

cellular regulatory processes. The automation of the sample acquisition and

subsequent analysis in high-throughput instruments that are capable of measuring

metabolites is posing a challenge on the necessary systematic storage and

computational processing of the experimental datasets. Whereas a multitude of

specialized software systems for individual instruments and preprocessing methods

exists, there is clearly a need for a free and platform-independent system that

allows the standardized and integrated storage and analysis of data obtained from

metabolomics experiments. Currently there exists no such system that on the one

hand supports preprocessing of raw datasets but also allows to visualize and

integrate the results of higher level statistical analyses within a functional

genomics context.

RESULTS: To facilitate the systematic storage, analysis and integration of

metabolomics experiments, we have implemented MeltDB, a web-based software

platform for the analysis and annotation of datasets from metabolomics

experiments. MeltDB supports open file formats (netCDF, mzXML, mzDATA) and

facilitates the integration and evaluation of existing preprocessing methods. The

system provides researchers with means to consistently describe and store their

experimental datasets. Comprehensive analysis and visualization features of

metabolomics datasets are offered to the community through a web-based user

interface. The system covers the process from raw data to the visualization of

results in a knowledge-based background and is integrated into the context of

existing software platforms of genomics and transcriptomics at Bielefeld

University. We demonstrate the potential of MeltDB by means of a sample

experiment where we dissect the influence of three different carbon sources on

the gram-negative bacterium Xanthomonas campestris pv. campestris on the level of

measured metabolites. Experimental data are stored, analyzed and annotated within

MeltDB and accessible via the public MeltDB web server.

AVAILABILITY: The system is publicly available at

http://meltdb.cebitec.uni-bielefeld.de.

DOI: 10.1093/bioinformatics/btn452

PMID: 18765459 [Indexed for MEDLINE]

2333. Immunogenetics. 2008 Dec;60(12):759-65. doi: 10.1007/s00251-008-0330-2. Epub 2008

Sep 3.

MHC motif viewer.

Rapin N(1), Hoof I, Lund O, Nielsen M.

Author information:

(1)Center for Biological Sequence Analysis, Department of Systems Biology,

Technical University of Denmark, Lyngby, Denmark.

In vertebrates, the major histocompatibility complex (MHC) presents peptides to

the immune system. In humans, MHCs are called human leukocyte antigens (HLAs),

and some of the loci encoding them are the most polymorphic in the human genome.

Different MHC molecules present different subsets of peptides, and knowledge of

their binding specificities is important for understanding the differences in the

immune response between individuals. Knowledge of motifs may be used to identify

epitopes, to understand the MHC restriction of epitopes, and to compare the

specificities of different MHC molecules. Algorithms that predict which peptides

MHC molecules bind have recently been developed and cover many different alleles,

but the utility of these algorithms is hampered by the lack of tools for browsing

and comparing the specificity of these molecules. We have, therefore, developed a

web server, MHC motif viewer, that allows the display of the likely binding motif

for all human class I proteins of the loci HLA A, B, C, and E and for MHC class I

molecules from chimpanzee (Pan troglodytes), rhesus monkey (Macaca mulatta), and

mouse (Mus musculus). Furthermore, it covers all HLA-DR protein sequences. A

special viewing feature, MHC fight, allows for display of the specificity of two

different MHC molecules side by side. We show how the web server can be used to

discover and display surprising similarities as well as differences between MHC

molecules within and between different species. The MHC motif viewer is available

at http://www.cbs.dtu.dk/biotools/MHCMotifViewer/ .

DOI: 10.1007/s00251-008-0330-2

PMCID: PMC2613509

PMID: 18766337 [Indexed for MEDLINE]

2334. J Comput Biol. 2008 Dec;15(10):1347-63. doi: 10.1089/cmb.2007.0176.

BayesMD: flexible biological modeling for motif discovery.

Tang MH(1), Krogh A, Winther O.

Author information:

(1)Bioinformatics Centre, Department of Molecular Biology, University of

Copenhagen, Copenhagen, Denmark.

We present BayesMD, a Bayesian Motif Discovery model with several new features.

Three different types of biological a priori knowledge are built into the

framework in a modular fashion. A mixture of Dirichlets is used as prior over

nucleotide probabilities in binding sites. It is trained on transcription factor

(TF) databases in order to extract the typical properties of TF binding sites. In

a similar fashion we train organism-specific priors for the background sequences.

Lastly, we use a prior over the position of binding sites. This prior represents

information complementary to the motif and background priors coming from

conservation, local sequence complexity, nucleosome occupancy, etc. and

assumptions about the number of occurrences. The Bayesian inference is carried

out using a combination of exact marginalization (multinomial parameters) and

sampling (over the position of sites). Robust sampling results are achieved using

the advanced sampling method parallel tempering. In a post-analysis step

candidate motifs with high marginal probability are found by searching among

those motifs that contain sites that occur frequently. Thereby, maximum a

posteriori inference for the motifs is avoided and the marginal probabilities can

be used directly to assess the significance of the findings. The framework is

benchmarked against other methods on a number of real and artificial data sets.

The accompanying prediction server, documentation, software, models and data are

available from http://bayesmd.binf.ku.dk/.

DOI: 10.1089/cmb.2007.0176

PMID: 19040368 [Indexed for MEDLINE]

2335. BMC Bioinformatics. 2008 Nov 28;9:503. doi: 10.1186/1471-2105-9-503.

ESLpred2: improved method for predicting subcellular localization of eukaryotic

proteins.

Garg A(1), Raghava GP.

Author information:

(1)Department of Biotechnology, Panjab University, Chandigarh, India.

aarti@imtech.res.in

BACKGROUND: The expansion of raw protein sequence databases in the post genomic

era and availability of fresh annotated sequences for major localizations

particularly motivated us to introduce a new improved version of our previously

forged eukaryotic subcellular localizations prediction method namely "ESLpred".

Since, subcellular localization of a protein offers essential clues about its

functioning, hence, availability of localization predictor would definitely aid

and expedite the protein deciphering studies. However, robustness of a predictor

is highly dependent on the superiority of dataset and extracted protein

attributes; hence, it becomes imperative to improve the performance of presently

available method using latest dataset and crucial input features.

RESULTS: Here, we describe augmentation in the prediction performance obtained

for our most popular ESLpred method using new crucial features as an input to

Support Vector Machine (SVM). In addition, recently available, highly

non-redundant dataset encompassing three kingdoms specific protein sequence sets;

1198 fungi sequences, 2597 from animal and 491 plant sequences were also included

in the present study. First, using the evolutionary information in the form of

profile composition along with whole and N-terminal sequence composition as an

input feature vector of 440 dimensions, overall accuracies of 72.7, 75.8 and

74.5% were achieved respectively after five-fold cross-validation. Further,

enhancement in performance was observed when similarity search based results were

coupled with whole and N-terminal sequence composition along with profile

composition by yielding overall accuracies of 75.9, 80.8, 76.6% respectively;

best accuracies reported till date on the same datasets.

CONCLUSION: These results provide confidence about the reliability and accurate

prediction of SVM modules generated in the present study using sequence and

profile compositions along with similarity search based results. The presently

developed modules are implemented as web server "ESLpred2" available at

http://www.imtech.res.in/raghava/eslpred2/.

DOI: 10.1186/1471-2105-9-503

PMCID: PMC2612013

PMID: 19038062 [Indexed for MEDLINE]

2336. Bioinformatics. 2008 Nov 15;24(22):2632-3. doi: 10.1093/bioinformatics/btn488.

Epub 2008 Sep 16.

TESE: generating specific protein structure test set ensembles.

Sirocco F(1), Tosatto SC.

Author information:

(1)Department of Biology, University of Padova, Viale G. Colombo 3, 35131 Padova,

Italy.

TESE is a web server for the generation of test sets of protein sequences and

structures fulfilling a number of different criteria. At least three different

use cases can be envisaged: (i) benchmarking of novel methods; (ii) test sets

tailored for special needs and (iii) extending available datasets. The CATH

structure classification is used to control structural/sequence redundancy and a

variety of structural quality parameters can be used to interactively select

protein subsets with specific characteristics, e.g. all X-ray structures of

alpha-helical repeat proteins with more than 120 residues and resolution <2.0 A.

The output includes FASTA-formatted sequences, PDB files and a clickable HTML

index file containing images of the selected proteins. Multiple subsets for

cross-validation are also supported.AVAILABILITY: The TESE server is available

for non-commercial use at URL: http://protein.bio.unipd.it/tese/.

DOI: 10.1093/bioinformatics/btn488

PMID: 18796478 [Indexed for MEDLINE]

2337. Biochem Biophys Res Commun. 2008 Nov 14;376(2):321-5. doi:

10.1016/j.bbrc.2008.08.125. Epub 2008 Sep 5.

ProtIdent: a web server for identifying proteases and their types by fusing

functional domain and sequential evolution information.

Chou KC(1), Shen HB.

Author information:

(1)Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong

University, 800 Dongchuan Road, Shanghai, 200240, China.

kcchou@gordonlifescience.org

Proteases are vitally important to life cycles and have become a main target in

drug development. According to their action mechanisms, proteases are classified

into six types: (1) aspartic, (2) cysteine, (3) glutamic, (4) metallo, (5)

serine, and (6) threonine. Given the sequence of an uncharacterized protein, can

we identify whether it is a protease or non-protease? If it is, what type does it

belong to? To address these problems, a 2-layer predictor, called "ProtIdent", is

developed by fusing the functional domain and sequential evolution information:

the first layer is for identifying the query protein as protease or non-protease;

if it is a protease, the process will automatically go to the second layer to

further identify it among the six types. The overall success rates in both cases

by rigorous cross-validation tests were higher than 92%. ProtIdent is freely

accessible to the public as a web server at

http://www.csbio.sjtu.edu.cn/bioinf/Protease.

DOI: 10.1016/j.bbrc.2008.08.125

PMID: 18774775 [Indexed for MEDLINE]

2338. BMC Genomics. 2008 Nov 4;9:525. doi: 10.1186/1471-2164-9-525.

NemaPath: online exploration of KEGG-based metabolic pathways for nematodes.

Wylie T(1), Martin J, Abubucker S, Yin Y, Messina D, Wang Z, McCarter JP, Mitreva

M.

Author information:

(1)The Genome Center at Washington University School of Medicine, St, Louis, MO

63108, USA. twylie@wustl.edu

BACKGROUND: Nematode.net http://www.nematode.net is a web-accessible resource for

investigating gene sequences from parasitic and free-living nematode genomes.

Beyond the well-characterized model nematode C. elegans, over 500,000 expressed

sequence tags (ESTs) and nearly 600,000 genome survey sequences (GSSs) have been

generated from 36 nematode species as part of the Parasitic Nematode Genomics

Program undertaken by the Genome Center at Washington University School of

Medicine. However, these sequencing data are not present in most publicly

available protein databases, which only include sequences in Swiss-Prot.

Swiss-Prot, in turn, relies on GenBank/Embl/DDJP for predicted proteins from

complete genomes or full-length proteins.

DESCRIPTION: Here we present the NemaPath pathway server, a web-based

pathway-level visualization tool for navigating putative metabolic pathways for

over 30 nematode species, including 27 parasites. The NemaPath approach consists

of two parts: 1) a backend tool to align and evaluate nematode genomic sequences

(curated EST contigs) against the annotated Kyoto Encyclopedia of Genes and

Genomes (KEGG) protein database; 2) a web viewing application that displays

annotated KEGG pathway maps based on desired confidence levels of primary

sequence similarity as defined by a user. NemaPath also provides cross-referenced

access to nematode genome information provided by other tools available on

Nematode.net, including: detailed NemaGene EST cluster information; putative

translations; GBrowse EST cluster views; links from nematode data to external

databases for corresponding synonymous C. elegans counterparts, subject matches

in KEGG's gene database, and also KEGG Ontology (KO) identification.

CONCLUSION: The NemaPath server hosts metabolic pathway mappings for 30 nematode

species and is available on the World Wide Web at

http://nematode.net/cgi-bin/keggview.cgi. The nematode source sequences used for

the metabolic pathway mappings are available via FTP

http://www.nematode.net/FTP/index.php, as provided by the Genome Center at

Washington University School of Medicine.

DOI: 10.1186/1471-2164-9-525

PMCID: PMC2588608

PMID: 18983679 [Indexed for MEDLINE]

2339. BMC Genomics. 2008 Nov 2;9:519. doi: 10.1186/1471-2164-9-519.

The prediction of protein-protein interaction networks in rice blast fungus.

He F(1), Zhang Y, Chen H, Zhang Z, Peng YL.

Author information:

(1)State Key Laboratory for ArgoBiotechnology, College of Biological Sciences,

China Agricultural University, Beijing 100193, PR China. plane332000@yahoo.com.cn

BACKGROUND: Protein-protein interaction (PPI) maps are useful tools for

investigating the cellular functions of genes. Thus far, large-scale PPI mapping

projects have not been implemented for the rice blast fungus Magnaporthe grisea,

which is responsible for the most severe rice disease. Inspired by recent

advances in PPI prediction, we constructed a PPI map of this important fungus.

RESULTS: Using a well-recognized interolog approach, we have predicted 11,674

interactions among 3,017 M. grisea proteins. Although the scale of the

constructed map covers approximately only one-fourth of the M. grisea's proteome,

it is the first PPI map for this crucial organism and will therefore provide new

insights into the functional genomics of the rice blast fungus. Focusing on the

network topology of proteins encoded by known pathogenicity genes, we have found

that pathogenicity proteins tend to interact with higher numbers of proteins. The

pathogenicity proteins and their interacting partners in the entire network were

then used to construct a subnet called a pathogenicity network. These data may

provide further clues for the study of these pathogenicity proteins. Finally, it

has been established that secreted proteins in M. grisea interact with fewer

proteins. These secreted proteins and their interacting partners were also

compiled into a network of secreted proteins, which may be helpful in

constructing an interactome between the rice blast fungus and rice.

CONCLUSION: We predicted the PPIs of M. grisea and compiled them into a database

server called MPID. It is hoped that MPID will provide new hints as to the

functional genomics of this fungus. MPID is available at

http://bioinformatics.cau.edu.cn/zzd\_lab/MPID.html.

DOI: 10.1186/1471-2164-9-519

PMCID: PMC2601049

PMID: 18976500 [Indexed for MEDLINE]

2340. Acta Crystallogr D Biol Crystallogr. 2008 Nov;64(Pt 11):1093-109. doi:

10.1107/S0907444908027388. Epub 2008 Oct 18.

PURY: a database of geometric restraints of hetero compounds for refinement in

complexes with macromolecular structures.

Andrejasic M(1), Praaenikar J, Turk D.

Author information:

(1)Department of Biochemistry and Molecular and Structural Biology, Joaeef Stefan

Institute, Jamova 39, 1000 Ljubljana, Slovenia.

The number and variety of macromolecular structures in complex with ;hetero'

ligands is growing. The need for rapid delivery of correct geometric parameters

for their refinement, which is often crucial for understanding the biological

relevance of the structure, is growing correspondingly. The current standard for

describing protein structures is the Engh-Huber parameter set. It is an expert

data set resulting from selection and analysis of the crystal structures gathered

in the Cambridge Structural Database (CSD). Clearly, such a manual approach

cannot be applied to the vast and ever-growing number of chemical compounds.

Therefore, a database, named PURY, of geometric parameters of chemical compounds

has been developed, together with a server that accesses it. PURY is a

compilation of the whole CSD. It contains lists of atom classes and bonds

connecting them, as well as angle, chirality, planarity and conformation

parameters. The current compilation is based on CSD 5.28 and contains 1978 atom

classes and 32,702 bonding, 237,068 angle, 201,860 dihedral and 64,193 improper

geometric restraints. Analysis has confirmed that the restraints from the PURY

database are suitable for use in macromolecular crystal structure refinement and

should be of value to the crystallographic community. The database can be

accessed through the web server http://pury.ijs.si/, which creates topology and

parameter files from deposited coordinates in suitable forms for the refinement

programs MAIN, CNS and REFMAC. In the near future, the server will move to the

CSD website http://pury.ccdc.cam.ac.uk/.

DOI: 10.1107/S0907444908027388

PMID: 19020347 [Indexed for MEDLINE]

2341. Ann Hum Genet. 2008 Nov;72(Pt 6):834-47. doi: 10.1111/j.1469-1809.2008.00469.x.

Epub 2008 Aug 13.

Efficient association study design via power-optimized tag SNP selection.

Han B(1), Kang HM, Seo MS, Zaitlen N, Eskin E.

Author information:

(1)Department of Computer Science and Engineering, University of California, San

Diego, La Jolla, CA 92093, USA.

Discovering statistical correlation between causal genetic variation and clinical

traits through association studies is an important method for identifying the

genetic basis of human diseases. Since fully resequencing a cohort is

prohibitively costly, genetic association studies take advantage of local

correlation structure (or linkage disequilibrium) between single nucleotide

polymorphisms (SNPs) by selecting a subset of SNPs to be genotyped (tag SNPs).

While many current association studies are performed using commercially available

high-throughput genotyping products that define a set of tag SNPs, choosing tag

SNPs remains an important problem for both custom follow-up studies as well as

designing the high-throughput genotyping products themselves. The most widely

used tag SNP selection method optimizes the correlation between SNPs (r(2)).

However, tag SNPs chosen based on an r(2) criterion do not necessarily maximize

the statistical power of an association study. We propose a study design

framework that chooses SNPs to maximize power and efficiently measures the power

through empirical simulation. Empirical results based on the HapMap data show

that our method gains considerable power over a widely used r(2)-based method, or

equivalently reduces the number of tag SNPs required to attain the desired power

of a study. Our power-optimized 100k whole genome tag set provides equivalent

power to the Affymetrix 500k chip for the CEU population. For the design of

custom follow-up studies, our method provides up to twice the power increase

using the same number of tag SNPs as r(2)-based methods. Our method is publicly

available via web server at http://design.cs.ucla.edu.

DOI: 10.1111/j.1469-1809.2008.00469.x

PMCID: PMC2574965

PMID: 18702637 [Indexed for MEDLINE]

2342. Behav Res Methods. 2008 Nov;40(4):1129-33. doi: 10.3758/BRM.40.4.1129.

The Creative task Creator: a tool for the generation of customized, Web-based

creativity tasks.

Pretz JE(1), Link JA.

Author information:

(1)Illinois Wesleyan University, Bloomington, Illinois 61701, USA. jpretz@iwu.edu

This article presents a Web-based tool for the creation of divergent-thinking and

open-ended creativity tasks. A Java program generates HTML forms with PHP

scripting that run an Alternate Uses Task and/or open-ended response items.

Researchers may specify their own instructions, objects, and time limits, or use

default settings. Participants can also be prompted to select their best

responses to the Alternate Uses Task (Silvia et al., 2008). Minimal programming

knowledge is required. The program runs on any server, and responses are recorded

in a standard MySQL database. Responses can be scored using the consensual

assessment technique (Amabile, 1996) or Torrance's (1998) traditional scoring

method. Adoption of this Web-based tool should facilitate creativity research

across cultures and access to eminent creators. The Creative Task Creator may be

downloaded from the Psychonomic Society's Archive of Norms, Stimuli, and Data,

www.psychonomic.org/archive.

DOI: 10.3758/BRM.40.4.1129

PMID: 19001404 [Indexed for MEDLINE]

2343. Bioinformatics. 2008 Nov 1;24(21):2539-41. doi: 10.1093/bioinformatics/btn466.

Epub 2008 Sep 1.

Bosque: integrated phylogenetic analysis software.

Ramírez-Flandes S(1), Ulloa O.

Author information:

(1)Departmento de Oceanografía, Universidad de Concepción, Casilla 160-C,

Concepción, Chile. sram@profc.udec.cl

Phylogenetic analyses today involve dealing with computer files in different

formats and often several computer programs. Although some widely used

applications have integrated important functionalities for such analyses, they

still work with local resources only: input/output files (users have to manage

them) and local computing (users have sometimes to leave their programs, on their

desktop computers, running for extended periods of time). To address these

problems we have developed 'Bosque', a multi-platform client-server software that

performs standard phylogenetic tasks either locally or remotely on servers, and

integrates the results on a local relational database. Bosque performs sequence

alignments and graphical visualization and editing of trees, thus providing a

powerful environment that integrates all the steps of phylogenetic

analyses.AVAILABILITY: http://bosque.udec.cl

DOI: 10.1093/bioinformatics/btn466

PMID: 18762483 [Indexed for MEDLINE]

2344. Bioinformatics. 2008 Nov 1;24(21):2526-33. doi: 10.1093/bioinformatics/btn459.

Epub 2008 Aug 26.

PathCase: pathways database system.

Elliott B(1), Kirac M, Cakmak A, Yavas G, Mayes S, Cheng E, Wang Y, Gupta C,

Ozsoyoglu G, Meral Ozsoyoglu Z.

Author information:

(1)Department of Electrical Engineering and Computer Science, Case Western

Reserve University, Cleveland, OH 44106, USA.

Comment in

Bioinformatics. 2009 Oct 15;25(20):2773.

MOTIVATION: As the blueprints of cellular actions, biological pathways

characterize the roles of genomic entities in various cellular mechanisms, and as

such, their availability, manipulation and queriability over the web is important

to facilitate ongoing biological research.

RESULTS: In this article, we present the new features of PathCase, a system to

store, query, visualize and analyze metabolic pathways at different levels of

genetic, molecular, biochemical and organismal detail. The new features include:

(i) a web-based system with a new architecture, containing a server-side and a

client-side, and promoting scalability, and flexible and easy adaptation of

different pathway databases, (ii) an interactive client-side visualization tool

for metabolic pathways, with powerful visualization capabilities, and with

integrated gene and organism viewers, (iii) two distinct querying capabilities:

an advanced querying interface for computer savvy users, and built-in queries for

ease of use, that can be issued directly from pathway visualizations and (iv) a

pathway functionality analysis tool. PathCase is now available for three

different datasets, namely, KEGG pathways data, sample pathways from the

literature and BioCyc pathways for humans.

AVAILABILITY: Available online at http://nashua.case.edu/pathways

DOI: 10.1093/bioinformatics/btn459

PMID: 18728044 [Indexed for MEDLINE]

2345. Eur Spine J. 2008 Nov;17(11):1497-506. doi: 10.1007/s00586-008-0779-6. Epub 2008

Oct 2.

Experience with an online prospective database on adolescent idiopathic

scoliosis: development and implementation.

Arlet V(1), Shilt J, Bersusky E, Abel M, Ouellet JA, Evans D, Menon KV, Kandziora

F, Shen F, Lamartina C, Adams M, Reddi V.

Author information:

(1)Department of Orthopedic Surgery, University of Virginia, Charlottesville, VA

22908, USA. va3e@virginia.edu

Considerable variability exists in the surgical treatment and outcomes of

adolescent idiopathic scoliosis (AIS). This is due to the lack of evidence-based

treatment guidelines and outcome measures. Although clinical trials have been

extolled as the highest form of evidence for evaluating treatment efficacy, the

disadvantage of cost, time, lack of feasibility, and ethical considerations

indicate a need for a new paradigm for evidence based research in this spinal

deformity. High quality clinical databases offer an alternative approach for

evidence-based research in medicine. So, we developed and established Scolisoft,

an international, multidimensional and relational database designed to be a

repository of surgical cases for AIS, and an active vehicle for standardized

surgical information in a format that would permit qualitative and quantitative

research and analysis. Here, we describe and discuss the utility of Scolisoft as

a new paradigm for evidence-based research on AIS. Scolisoft was developed using

dot.net platform and SQL server from Microsoft. All data is deidentified to

protect patient privacy. Scolisoft can be accessed at (www.scolisoft.org).

Collection of high quality data on surgical cases of AIS is a priority and

processes continue to improve the database quality. The database currently has 67

registered users from 21 countries. To date, Scolisoft has 200 detailed surgical

cases with pre, post, and follow up data. Scolisoft provides a structured process

and practical information for surgeons to benchmark their treatment methods

against other like treatments. Scolisoft is multifaceted and its use extends to

education of health care providers in training, patients, ability to mine

important data to stimulate research and quality improvement initiatives of

healthcare organizations.

DOI: 10.1007/s00586-008-0779-6

PMCID: PMC2583197

PMID: 18830720 [Indexed for MEDLINE]

2346. J Comput Biol. 2008 Nov;15(9):1187-94. doi: 10.1089/cmb.2008.0125.

Significance of gapped sequence alignments.

Newberg LA(1).

Author information:

(1)Center for Bioinformatics, Wadsworth Center, New York State Department of

Health, Albany, New York 12201-0509, USA.

Measurement of the the statistical significance of extreme sequence alignment

scores is key to many important applications, but it is difficult. To precisely

approximate alignment score significance, we draw random samples directly from a

well chosen, importance-sampling probability distribution. We apply our technique

to pairwise local sequence alignment of nucleic acid and amino acid sequences of

length up to 1000. For instance, using a BLOSUM62 scoring system for local

sequence alignment, we compute that the p-value of a score of 6000 for the

alignment of two sequences of length 1000 is (3.4 +/- 0.3) x 10(-1314). Further,

we show that the extreme value significance statistic for the local alignment

model that we examine does not follow a Gumbel distribution. A web server for

this application is available at

http://bayesweb.wadsworth.org/alignmentSignificanceV1/.

DOI: 10.1089/cmb.2008.0125

PMCID: PMC2737730

PMID: 18973434 [Indexed for MEDLINE]

2347. J Mol Recognit. 2008 Nov-Dec;21(6):431-41. doi: 10.1002/jmr.910.

Estimation and extraction of B-cell linear epitopes predicted by mathematical

morphology approaches.

Chang HT(1), Liu CH, Pai TW.

Author information:

(1)Graduate Institute of Molecular Systems Biomedicine, China Medical University,

Taichung, Taiwan, ROC.

B-cell epitope prediction facilitates the design and synthesis of short peptides

for various immunological applications. Several algorithms have been developed to

predict B-cell linear epitopes (LEs) from primary sequences of antigens,

providing important information for immunobiological experiments and antibody

design. This paper describes two robust methods, LE prediction with/without local

peak extraction (LEP-LP and LEP-NLP), based on antigenicity scale and

mathematical morphology for the prediction of B-cell LEs. Previous studies

revealed that LEs could occur in regions with low-to-moderate but not globally

high antigenicity scales. Hence, we developed a method adopting mathematical

morphology to extract local peaks from a linear combination of the propensity

scales of physico-chemical characteristics at each antigen residue. Comparison

among LEP-LP/LEP-NLP, BepiPred and BEPITOPE revealed that our algorithms

performed better in retrieving epitopes with low-to-moderate antigenicity and

achieved comparable performance according to receiver operation characteristics

(ROC) curve analysis. Of the identified LEs, over 30% were unable to be predicted

by BepiPred and BEPITOPE employing an average threshold of antigenicity index or

default settings. Our LEP-LP method provides a bioinformatics approach for

predicting B-cell LEs with low- to-moderate antigenicity. The web-based server

was established at http://biotools.cs.ntou.edu.tw/lepd\_antigenicity. php for free

use.

Copyright 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/jmr.910

PMID: 18680207 [Indexed for MEDLINE]

2348. PLoS Comput Biol. 2008 Nov;4(11):e1000213. doi: 10.1371/journal.pcbi.1000213.

Epub 2008 Nov 7.

Transmembrane topology and signal peptide prediction using dynamic bayesian

networks.

Reynolds SM(1), Käll L, Riffle ME, Bilmes JA, Noble WS.

Author information:

(1)Department of Electrical Engineering, University of Washington, Seattle,

Washington, United States of America.

Hidden Markov models (HMMs) have been successfully applied to the tasks of

transmembrane protein topology prediction and signal peptide prediction. In this

paper we expand upon this work by making use of the more powerful class of

dynamic Bayesian networks (DBNs). Our model, Philius, is inspired by a previously

published HMM, Phobius, and combines a signal peptide submodel with a

transmembrane submodel. We introduce a two-stage DBN decoder that combines the

power of posterior decoding with the grammar constraints of Viterbi-style

decoding. Philius also provides protein type, segment, and topology confidence

metrics to aid in the interpretation of the predictions. We report a relative

improvement of 13% over Phobius in full-topology prediction accuracy on

transmembrane proteins, and a sensitivity and specificity of 0.96 in detecting

signal peptides. We also show that our confidence metrics correlate well with the

observed precision. In addition, we have made predictions on all 6.3 million

proteins in the Yeast Resource Center (YRC) database. This large-scale study

provides an overall picture of the relative numbers of proteins that include a

signal-peptide and/or one or more transmembrane segments as well as a valuable

resource for the scientific community. All DBNs are implemented using the

Graphical Models Toolkit. Source code for the models described here is available

at http://noble.gs.washington.edu/proj/philius. A Philius Web server is available

at http://www.yeastrc.org/philius, and the predictions on the YRC database are

available at http://www.yeastrc.org/pdr.

DOI: 10.1371/journal.pcbi.1000213

PMCID: PMC2570248

PMID: 18989393 [Indexed for MEDLINE]

2349. BMC Bioinformatics. 2008 Oct 28;9:459. doi: 10.1186/1471-2105-9-459.

MetaMine--a tool to detect and analyse gene patterns in their environmental

context.

Bohnebeck U(1), Lombardot T, Kottmann R, Glöckner FO.

Author information:

(1)ttz Bremerhaven, Bremerhaven, Germany. bohnebeck@ttz-bremerhaven.de

BACKGROUND: Modern sequencing technologies allow rapid sequencing and

bioinformatic analysis of genomes and metagenomes. With every new sequencing

project a vast number of new proteins become available with many genes remaining

functionally unclassified based on evidences from sequence similarities alone.

Extending similarity searches with gene pattern approaches, defined as genes

sharing a distinct genomic neighbourhood, have shown to significantly improve the

number of functional assignments. Further functional evidences can be gained by

correlating these gene patterns with prevailing environmental parameters.

MetaMine was developed to approach the large pool of unclassified proteins by

searching for recurrent gene patterns across habitats based on key genes.

RESULTS: MetaMine is an interactive data mining tool which enables the detection

of gene patterns in an environmental context. The gene pattern search starts with

a user defined environmentally interesting key gene. With this gene a BLAST

search is carried out against the Microbial Ecological Genomics DataBase (MEGDB)

containing marine genomic and metagenomic sequences. This is followed by the

determination of all neighbouring genes within a given distance and a search for

functionally equivalent genes. In the final step a set of common genes present in

a defined number of distinct genomes is determined. The gene patterns found are

associated with their individual pattern instances describing gene order and

directions. They are presented together with information about the sample and the

habitat. MetaMine is implemented in Java and provided as a client/server

application with a user-friendly graphical user interface. The system was

evaluated with environmentally relevant genes related to the methane-cycle and

carbon monoxide oxidation.

CONCLUSION: MetaMine offers a targeted, semi-automatic search for gene patterns

based on expert input. The graphical user interface of MetaMine provides a

user-friendly overview of the computed gene patterns for further inspection in an

ecological context. Prevailing biological processes associated with a key gene

can be used to infer new annotations and shape hypotheses to guide further

analyses. The use-cases demonstrate that meaningful gene patterns can be quickly

detected using MetaMine.MetaMine is freely available for academic use from

http://www.megx.net/metamine.

DOI: 10.1186/1471-2105-9-459

PMCID: PMC2615450

PMID: 18957118 [Indexed for MEDLINE]

2350. BMC Genomics. 2008 Oct 27;9:505. doi: 10.1186/1471-2164-9-505.

RAId\_DbS: mass-spectrometry based peptide identification web server with

knowledge integration.

Alves G(1), Ogurtsov AY, Yu YK.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

NIH, Bethesda, MD 20894, USA. alves@ncbi.nlm.nih.gov

BACKGROUND: Existing scientific literature is a rich source of biological

information such as disease markers. Integration of this information with data

analysis may help researchers to identify possible controversies and to form

useful hypotheses for further validations. In the context of proteomics studies,

individualized proteomics era may be approached through consideration of amino

acid substitutions/modifications as well as information from disease studies.

Integration of such information with peptide searches facilitates speedy, dynamic

information retrieval that may significantly benefit clinical laboratory studies.

DESCRIPTION: We have integrated from various sources annotated single amino acid

polymorphisms, post-translational modifications, and their documented disease

associations (if they exist) into one enhanced database per organism. We have

also augmented our peptide identification software RAId\_DbS to take into account

this information while analyzing a tandem mass spectrum. In principle, one may

choose to respect or ignore the correlation of amino acid

polymorphisms/modifications within each protein. The former leads to targeted

searches and avoids scoring of unnecessary polymorphism/modification

combinations; the latter explores possible polymorphisms in a controlled fashion.

To facilitate new discoveries, RAId\_DbS also allows users to conduct searches

permitting novel polymorphisms as well as to search a knowledge database created

by the users.

CONCLUSION: We have finished constructing enhanced databases for 17 organisms.

The web link to RAId\_DbS and the enhanced databases is

http://www.ncbi.nlm.nih.gov/CBBResearch/qmbp/RAId\_DbS/index.html. The relevant

databases and binaries of RAId\_DbS for Linux, Windows, and Mac OS X are available

for download from the same web page.

DOI: 10.1186/1471-2164-9-505

PMCID: PMC2605478

PMID: 18954448 [Indexed for MEDLINE]

2351. J Environ Qual. 2008 Oct 23;37(6):2392-6. doi: 10.2134/jeq2007.0536. Print 2008

Nov-Dec.

CProb: a computational tool for conducting conditional probability analysis.

Hollister JW(1), Walker HA, Paul JF.

Author information:

(1)U.S. Environmental Protection Agency, Office of Research and Development,

National Health and Environmental Effects Research Lab., Atlantic Ecology Div.,

27 Tarzwell Drive, Narragansett, RI 02882, USA. Hollister.jeff@epa.gov

Conditional probability is the probability of observing one event given that

another event has occurred. In an environmental context, conditional probability

helps to assess the association between an environmental contaminant (i.e., the

stressor) and the ecological condition of a resource (i.e., the response). These

analyses, when combined with controlled experiments and other methodologies, show

great promise in evaluating ecological conditions from observational data and in

defining water quality and other environmental criteria. Current applications of

conditional probability analysis (CPA) are largely done via scripts or cumbersome

spreadsheet routines, which may prove daunting to end-users and do not provide

access to the underlying scripts. Combining spreadsheets with scripts eases

computation through a familiar interface (i.e., Microsoft Excel) and creates a

transparent process through full accessibility to the scripts. With this in mind,

we developed a software application, CProb, as an Add-in for Microsoft Excel with

R, R(D)com Server, and Visual Basic for Applications. CProb calculates and plots

scatterplots, empirical cumulative distribution functions, and conditional

probability. In this short communication, we describe CPA, our motivation for

developing a CPA tool, and our implementation of CPA as a Microsoft Excel Add-in.

Further, we illustrate the use of our software with two examples: a water quality

example and a landscape example. CProb is freely available for download at

http://www.epa.gov/emap/nca/html/regions/cprob.

DOI: 10.2134/jeq2007.0536

PMID: 18948494 [Indexed for MEDLINE]

2352. BMC Microbiol. 2008 Oct 20;8:185. doi: 10.1186/1471-2180-8-185.

PrimerSNP: a web tool for whole-genome selection of allele-specific and common

primers of phylogenetically-related bacterial genomic sequences.

Yao J(1), Lin H, Van Deynze A, Doddapaneni H, Francis M, Lemos EG, Civerolo EL.

Author information:

(1)Citrus Research Board, 323 W. Oak Street, Visalia, CA 93291, USA.

jiqiangyao@yahoo.com

BACKGROUND: The increasing number of genomic sequences of bacteria makes it

possible to select unique SNPs of a particular strain/species at the whole genome

level and thus design specific primers based on the SNPs. The high similarity of

genomic sequences among phylogenetically-related bacteria requires the

identification of the few loci in the genome that can serve as unique markers for

strain differentiation. PrimerSNP attempts to identify reliable strain-specific

markers, on which specific primers are designed for pathogen detection purpose.

RESULTS: PrimerSNP is an online tool to design primers based on strain specific

SNPs for multiple strains/species of microorganisms at the whole genome level.

The allele-specific primers could distinguish query sequences of one strain from

other homologous sequences by standard PCR reaction. Additionally, PrimerSNP

provides a feature for designing common primers that can amplify all the

homologous sequences of multiple strains/species of microorganisms. PrimerSNP is

freely available at http://cropdisease.ars.usda.gov/~primer.

CONCLUSION: PrimerSNP is a high-throughput specific primer generation tool for

the differentiation of phylogenetically-related strains/species. Experimental

validation showed that this software had a successful prediction rate of 80.4 -

100% for strain specific primer design.

DOI: 10.1186/1471-2180-8-185

PMCID: PMC2579435

PMID: 18937861 [Indexed for MEDLINE]

2353. BMC Bioinformatics. 2008 Oct 10;9:429. doi: 10.1186/1471-2105-9-429.

Pairwise covariance adds little to secondary structure prediction but improves

the prediction of non-canonical local structure.

Bystroff C(1), Webb-Robertson BJ.

Author information:

(1)Department of Biology, Center for Biotechnology and Interdisciplinary Studies,

Rensselaer Polytechnic Institute, Troy NY, USA. bystrc@rpi.edu

BACKGROUND: Amino acid sequence probability distributions, or profiles, have been

used successfully to predict secondary structure and local structure in proteins.

Profile models assume the statistical independence of each position in the

sequence, but the energetics of protein folding is better captured in a scoring

function that is based on pairwise interactions, like a force field.

RESULTS: I-sites motifs are short sequence/structure motifs that populate the

protein structure database due to energy-driven convergent evolution. Here we

show that a pairwise covariant sequence model does not predict alpha helix or

beta strand significantly better overall than a profile-based model, but it does

improve the prediction of certain loop motifs. The finding is best explained by

considering secondary structure profiles as multivariant, all-or-none models,

which subsume covariant models. Pairwise covariance is nonetheless present and

energetically rational. Examples of negative design are present, where the

covariances disfavor non-native structures.

CONCLUSION: Measured pairwise covariances are shown to be statistically robust in

cross-validation tests, as long as the amino acid alphabet is reduced to nine

classes. An updated I-sites local structure motif library that provides sequence

covariance information for all types of local structure in globular proteins and

a web server for local structure prediction are available at

http://www.bioinfo.rpi.edu/bystrc/hmmstr/server.php.

DOI: 10.1186/1471-2105-9-429

PMCID: PMC2579440

PMID: 18847485 [Indexed for MEDLINE]

2354. Amino Acids. 2008 Oct;35(3):599-605. doi: 10.1007/s00726-008-0085-y. Epub 2008

Apr 19.

DPROT: prediction of disordered proteins using evolutionary information.

Sethi D(1), Garg A, Raghava GP.

Author information:

(1)Scientist and Head Bioinformatics Centre, Institute of Microbial Technology,

Sector 39A, Chandigarh, India.

The association of structurally disordered proteins with a number of diseases has

engendered enormous interest and therefore demands a prediction method that would

facilitate their expeditious study at molecular level. The present study

describes the development of a computational method for predicting disordered

proteins using sequence and profile compositions as input features for the

training of SVM models. First, we developed the amino acid and dipeptide

compositions based SVM modules which yielded sensitivities of 75.6 and 73.2%

along with Matthew's Correlation Coefficient (MCC) values of 0.75 and 0.60,

respectively. In addition, the use of predicted secondary structure content

(coil, sheet and helices) in the form of composition values attained a

sensitivity of 76.8% and MCC value of 0.77. Finally, the training of SVM models

using evolutionary information hidden in the multiple sequence alignment profile

improved the prediction performance by achieving a sensitivity value of 78% and

MCC of 0.78. Furthermore, when evaluated on an independent dataset of partially

disordered proteins, the same SVM module provided a correct prediction rate of

86.6%. Based on the above study, a web server ("DPROT") was developed for the

prediction of disordered proteins, which is available at

http://www.imtech.res.in/raghava/dprot/.

DOI: 10.1007/s00726-008-0085-y

PMID: 18425404 [Indexed for MEDLINE]

2355. Bioinformatics. 2008 Oct 1;24(19):2263-4. doi: 10.1093/bioinformatics/btn417.

Epub 2008 Aug 11.

Arabidopsis thaliana regulatory element analyzer.

Choudhury A(1), Lahiri A.

Author information:

(1)Department of Biophysics, Molecular Biology and Genetics, University of

Calcutta, 92 APC Road, Kolkata 700009, India. ananyo.c@gmail.com

In the Arabidopsis thaliana regulatory element analyzer (AtREA) server, we have

integrated sequence data, genome-wide expression data and functional annotation

data in three application modules which will be useful to identify major

regulatory targets of a user-provided cis-regulatory element (CRE), study

different features of CRE distribution and evaluate the role of a set of CREs in

the regulation of gene expression--independently as well as in combination with

other user-provided CREs.AVAILABILITY: AtREA is freely available at

http://www.bioinformatics.org/grn/atrea.html.

DOI: 10.1093/bioinformatics/btn417

PMID: 18694893 [Indexed for MEDLINE]

2356. Environ Microbiol. 2008 Oct;10(10):2894-8. doi: 10.1111/j.1462-2920.2008.01706.x.

Epub 2008 Jul 21.

probeCheck--a central resource for evaluating oligonucleotide probe coverage and

specificity.

Loy A(1), Arnold R, Tischler P, Rattei T, Wagner M, Horn M.

Author information:

(1)Department of Microbial Ecology, Universität Wien, Althanstrasse 14, A-1090

Wien, Austria.

The web server probeCheck, freely accessible at

http://www.microbial-ecology.net/probecheck, provides a pivotal forum for rapid

specificity and coverage evaluations of probes and primers against selected

databases of phylogenetic and functional marker genes. Currently, 24 widely used

sequence collections including the Ribosomal Database Project (RDP) II,

Greengenes, SILVA and the Functional Gene Pipeline/Repository can be queried. For

this purpose, probeCheck integrates a new online version of the popular ARB probe

match tool with free energy (DeltaG) calculations for each perfectly matched and

mismatched probe-target hybrid, allowing assessment of the theoretical binding

stabilities of oligo-target and non-target hybrids. For each output sequence, the

accession number, the GenBank taxonomy and a link to the respective entry at

GenBank, EMBL and, if applicable, the query database are displayed. Filtering

options allow customizing results on the output page. In addition, probeCheck is

linked with probe match tools of RDP II and Greengenes, NCBI blast, the

Oligonucleotide Properties Calculator, the two-state folding tool of the DINAMelt

server and the rRNA-targeted probe database probeBase. Taken together, these

features provide a multifunctional platform with maximal flexibility for the user

in the choice of databases and options for the evaluation of published and newly

developed probes and primers.

DOI: 10.1111/j.1462-2920.2008.01706.x

PMCID: PMC2613240

PMID: 18647333 [Indexed for MEDLINE]

2357. Gene. 2008 Oct 1;422(1-2):14-21. doi: 10.1016/j.gene.2008.06.014. Epub 2008 Jun

14.

Using protein binding site prediction to improve protein docking.

Huang B(1), Schroeder M.

Author information:

(1)Bioinformatics Group, Biotechnological Center, Technical University Dresden,

Germany. bhuang@biotec.tu-dresden.de

Predicting protein interaction interfaces and protein complexes are two important

related problems. For interface prediction, there are a number of tools, such as

PPI-Pred, PPISP, PINUP, Promate, and SPPIDER, which predict enzyme-inhibitor

interfaces with success rates of 23% to 55% and other interfaces with 10% to 28%

on a benchmark dataset of 62 complexes. Here, we develop, metaPPI, a meta server

for interface prediction. It significantly improves prediction success rates to

70% for enzyme-inhibitor and 44% for other interfaces. As shown with Promate,

predicted interfaces can be used to improve protein docking. Here, we follow this

idea using the meta server instead of individual predictions. We confirm that

filtering with predicted interfaces significantly improves candidate generation

in rigid-body docking based on shape complementarity. Finally, we show that the

initial ranking of candidate solutions in rigid-body docking can be further

improved for the class of enzyme-inhibitor complexes by a geometrical scoring

which rewards deep pockets. A web server of metaPPI is available at

scoppi.tu-dresden.de/metappi. The source code of our docking algorithm BDOCK is

also available at www.biotec.tu-dresden.de /approximately bhuang/bdock.

DOI: 10.1016/j.gene.2008.06.014

PMID: 18616991 [Indexed for MEDLINE]

2358. Proteins. 2008 Oct;73(1):72-86. doi: 10.1002/prot.22052.

MolAxis: efficient and accurate identification of channels in macromolecules.

Yaffe E(1), Fishelovitch D, Wolfson HJ, Halperin D, Nussinov R.

Author information:

(1)School of Computer Science, Raymond and Beverly Sackler Faculty for Exact

Science, Tel Aviv University, Tel Aviv 69978, Israel.

Channels and cavities play important roles in macromolecular functions, serving

as access/exit routes for substrates/products, cofactor and drug binding,

catalytic sites, and ligand/protein. In addition, channels formed by

transmembrane (TM) proteins serve as transporters and ion channels. MolAxis is a

new sensitive and fast tool for the identification and classification of channels

and cavities of various sizes and shapes in macromolecules. MolAxis constructs

corridors, which are pathways that represent probable routes taken by small

molecules passing through channels. The outer medial axis of the molecule is the

collection of points that have more than one closest atom. It is composed of

two-dimensional surface patches and can be seen as a skeleton of the complement

of the molecule. We have implemented in MolAxis a novel algorithm that uses

state-of-the-art computational geometry techniques to approximate and scan a

useful subset of the outer medial axis, thereby reducing the dimension of the

problem and consequently rendering the algorithm extremely efficient. MolAxis is

designed to identify channels that connect buried cavities to the outside of

macromolecules and to identify TM channels in proteins. We apply MolAxis to

enzyme cavities and TM proteins. We further utilize MolAxis to monitor channel

dimensions along Molecular Dynamics trajectories of a human Cytochrome P450.

MolAxis constructs high quality corridors for snapshots at picosecond time-scale

intervals substantiating the gating mechanism in the 2e substrate access channel.

We compare our results with previous tools in terms of accuracy, performance and

underlying theoretical guarantees of finding the desired pathways. MolAxis is

available on line as a web-server and as a stand alone easy-to-use program

(http://bioinfo3d.cs.tau.ac.il/MolAxis/).

DOI: 10.1002/prot.22052

PMCID: PMC2693897

PMID: 18393395 [Indexed for MEDLINE]

2359. BMC Bioinformatics. 2008 Sep 19;9:386. doi: 10.1186/1471-2105-9-386.

The metagenomics RAST server - a public resource for the automatic phylogenetic

and functional analysis of metagenomes.

Meyer F(1), Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T,

Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA.

Author information:

(1)Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S

Cass Avenue, Argonne, IL 60439, USA. folker@mcs.anl.gov

BACKGROUND: Random community genomes (metagenomes) are now commonly used to study

microbes in different environments. Over the past few years, the major challenge

associated with metagenomics shifted from generating to analyzing sequences.

High-throughput, low-cost next-generation sequencing has provided access to

metagenomics to a wide range of researchers.

RESULTS: A high-throughput pipeline has been constructed to provide

high-performance computing to all researchers interested in using metagenomics.

The pipeline produces automated functional assignments of sequences in the

metagenome by comparing both protein and nucleotide databases. Phylogenetic and

functional summaries of the metagenomes are generated, and tools for comparative

metagenomics are incorporated into the standard views. User access is controlled

to ensure data privacy, but the collaborative environment underpinning the

service provides a framework for sharing datasets between multiple users. In the

metagenomics RAST, all users retain full control of their data, and everything is

available for download in a variety of formats.

CONCLUSION: The open-source metagenomics RAST service provides a new paradigm for

the annotation and analysis of metagenomes. With built-in support for multiple

data sources and a back end that houses abstract data types, the metagenomics

RAST is stable, extensible, and freely available to all researchers. This service

has removed one of the primary bottlenecks in metagenome sequence analysis - the

availability of high-performance computing for annotating the data.

http://metagenomics.nmpdr.org.

DOI: 10.1186/1471-2105-9-386

PMCID: PMC2563014

PMID: 18803844 [Indexed for MEDLINE]

2360. BMC Genomics. 2008 Sep 18;9:422. doi: 10.1186/1471-2164-9-422.

WebScipio: an online tool for the determination of gene structures using protein

sequences.

Odronitz F(1), Pillmann H, Keller O, Waack S, Kollmar M.

Author information:

(1)Max-Planck-Institut für Biophysikalische Chemie, Abteilung NMR-basierte

Strukturbiologie, Am Fassberg 11, 37077 Göttingen, Germany.

flod@nmr.mpibpc.mpg.de

BACKGROUND: Obtaining the gene structure for a given protein encoding gene is an

important step in many analyses. A software suited for this task should be

readily accessible, accurate, easy to handle and should provide the user with a

coherent representation of the most probable gene structure. It should be

rigorous enough to optimise features on the level of single bases and at the same

time flexible enough to allow for cross-species searches.

RESULTS: WebScipio, a web interface to the Scipio software, allows a user to

obtain the corresponding coding sequence structure of a here given a query

protein sequence that belongs to an already assembled eukaryotic genome. The

resulting gene structure is presented in various human readable formats like a

schematic representation, and a detailed alignment of the query and the target

sequence highlighting any discrepancies. WebScipio can also be used to identify

and characterise the gene structures of homologs in related organisms. In

addition, it offers a web service for integration with other programs.

CONCLUSION: WebScipio is a tool that allows users to get a high-quality gene

structure prediction from a protein query. It offers more than 250 eukaryotic

genomes that can be searched and produces predictions that are close to what can

be achieved by manual annotation, for in-species and cross-species searches

alike. WebScipio is freely accessible at http://www.webscipio.org.

DOI: 10.1186/1471-2164-9-422

PMCID: PMC2644328

PMID: 18801164 [Indexed for MEDLINE]

2361. Biol Direct. 2008 Sep 16;3:38. doi: 10.1186/1745-6150-3-38.

CAIcal: a combined set of tools to assess codon usage adaptation.

Puigbò P(1), Bravo IG, Garcia-Vallve S.

Author information:

(1)Department of Biochemistry and Biotechnology, Rovira i Virgili University

(URV), Campus Sescelades, c/Marcelli Domingo s/n, 43007 Tarragona, Spain.

puigboap@ncbi.nlm.nih.gov

BACKGROUND: The Codon Adaptation Index (CAI) was first developed to measure the

synonymous codon usage bias for a DNA or RNA sequence. The CAI quantifies the

similarity between the synonymous codon usage of a gene and the synonymous codon

frequency of a reference set.

RESULTS: We describe here CAIcal, a web-server available at

http://genomes.urv.es/CAIcal that includes a complete set of utilities related

with the CAI. The server provides useful important features, such as the

calculation and graphical representation of the CAI along either an individual

sequence or a protein multiple sequence alignment translated to DNA. The

automated calculation of CAI and its expected value is also included as one of

the CAIcal tools. The software is also free to be downloaded as a stand alone

application for local use.

CONCLUSION: The CAIcal server provides a complete set of tools to assess codon

usage adaptation and to help in genome annotation.

DOI: 10.1186/1745-6150-3-38

PMCID: PMC2553769

PMID: 18796141 [Indexed for MEDLINE]

2362. Bioinformatics. 2008 Sep 15;24(18):2101-2. doi: 10.1093/bioinformatics/btn392.

Epub 2008 Jul 28.

HELIQUEST: a web server to screen sequences with specific alpha-helical

properties.

Gautier R(1), Douguet D, Antonny B, Drin G.

Author information:

(1)Institut de Pharmacologie Moléculaire et Cellulaire, Université de Nice Sophia

Antipolis and CNRS, Valbonne, France. gautier@ipmc.cnrs.fr

SUMMARY: HELIQUEST calculates the physicochemical properties and amino acid

composition of an alpha-helix and screens databank to identify protein segments

possessing similar features. This server is also dedicated to mutating helices

manually or automatically by genetic algorithm to design analogues of defined

features.

AVAILABILITY: http://heliquest.ipmc.cnrs.fr.

DOI: 10.1093/bioinformatics/btn392

PMID: 18662927 [Indexed for MEDLINE]

2363. Bioinformatics. 2008 Sep 15;24(18):2079-85. doi: 10.1093/bioinformatics/btn378.

Epub 2008 Jul 19.

Data-driven extraction of relative reasoning rules to limit combinatorial

explosion in biodegradation pathway prediction.

Fenner K(1), Gao J, Kramer S, Ellis L, Wackett L.

Author information:

(1)Eawag, Swiss Federal Institute of Aquatic Science and Technology, CH-8600

Dübendorf, Switzerland. kathrin.fenner@eawag.ch

MOTIVATION: The University of Minnesota Pathway Prediction System (UM-PPS) is a

rule-based expert system to predict plausible biodegradation pathways for organic

compounds. However, iterative application of these rules to generate

biodegradation pathways leads to combinatorial explosion. We use data from known

biotransformation pathways to rationally determine biotransformation priorities

(relative reasoning rules) to limit this explosion.

RESULTS: A total of 112 relative reasoning rules were identified and implemented.

In one prediction step, i.e. as per one generation predicted, the use of relative

reasoning decreases the predicted biotransformations by over 25% for 50 compounds

used to generate the rules and by about 15% for an external validation set of 47

xenobiotics, including pesticides, biocides and pharmaceuticals. The percentage

of correctly predicted, experimentally known products remains at 75% when

relative reasoning is used. The set of relative reasoning rules identified,

therefore, effectively reduces the number of predicted transformation products

without compromising the quality of the predictions.

AVAILABILITY: The UM-PPS server is freely available on the web to all users at

the time of submission of this manuscript and will be available following

publication at http://umbbd.msi.umn.edu/predict/.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btn378

PMID: 18641402 [Indexed for MEDLINE]

2364. Bioinformatics. 2008 Sep 15;24(18):2110-1. doi: 10.1093/bioinformatics/btn363.

Epub 2008 Jul 16.

The protein information and property explorer: an easy-to-use, rich-client web

application for the management and functional analysis of proteomic data.

Ramos H(1), Shannon P, Aebersold R.

Author information:

(1)Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103-8904,

USA.

MOTIVATION: Mass spectrometry experiments in the field of proteomics produce

lists containing tens to thousands of identified proteins. With the protein

information and property explorer (PIPE), the biologist can acquire functional

annotations for these proteins and explore the enrichment of the list, or

fraction thereof, with respect to functional classes. These protein lists may be

saved for access at a later time or different location. The PIPE is interoperable

with the Firegoose and the Gaggle, permitting wide-ranging data exploration and

analysis. The PIPE is a rich-client web application which uses AJAX capabilities

provided by the Google Web Toolkit, and server-side data storage using Hibernate.

AVAILABILITY: http://pipe.systemsbiology.net.

DOI: 10.1093/bioinformatics/btn363

PMCID: PMC2638980

PMID: 18635572 [Indexed for MEDLINE]

2365. Bioinformatics. 2008 Sep 15;24(18):2094-5. doi: 10.1093/bioinformatics/btn371.

Epub 2008 Jul 16.

MetalDetector: a web server for predicting metal-binding sites and disulfide

bridges in proteins from sequence.

Lippi M(1), Passerini A, Punta M, Rost B, Frasconi P.

Author information:

(1)Dipartimento di Sistemi e Informatica, Machine Learning and Neural Networks

Group, Università degli Studi di Firenze, Via di Santa Marta 3, 50139 Firenze,

Italy.

The web server MetalDetector classifies histidine residues in proteins into one

of two states (free or metal bound) and cysteines into one of three states (free,

metal bound or disulfide bridged). A decision tree integrates predictions from

two previously developed methods (DISULFIND and Metal Ligand Predictor).

Cross-validated performance assessment indicates that our server predicts

disulfide bonding state at 88.6% precision and 85.1% recall, while it identifies

cysteines and histidines in transition metal-binding sites at 79.9% precision and

76.8% recall, and at 60.8% precision and 40.7% recall, respectively.AVAILABILITY:

Freely available at http://metaldetector.dsi.unifi.it.

SUPPLEMENTARY INFORMATION: Details and data can be found at

http://metaldetector.dsi.unifi.it/help.php.

DOI: 10.1093/bioinformatics/btn371

PMCID: PMC2732205

PMID: 18635571 [Indexed for MEDLINE]

2366. Bioinformatics. 2008 Sep 15;24(18):2002-9. doi: 10.1093/bioinformatics/btn353.

Epub 2008 Jul 16.

Accurate prediction of stability changes in protein mutants by combining machine

learning with structure based computational mutagenesis.

Masso M(1), Vaisman II.

Author information:

(1)Department of Bioinformatics and Computational Biology, Laboratory for

Structural Bioinformatics, George Mason University, 10900 University Blvd, MSN

5B3, Manassas, VA 20110, USA.

MOTIVATION: Accurate predictive models for the impact of single amino acid

substitutions on protein stability provide insight into protein structure and

function. Such models are also valuable for the design and engineering of new

proteins. Previously described methods have utilized properties of protein

sequence or structure to predict the free energy change of mutants due to thermal

(DeltaDeltaG) and denaturant (DeltaDeltaG(H2O)) denaturations, as well as mutant

thermal stability (DeltaT(m)), through the application of either computational

energy-based approaches or machine learning techniques. However, accuracy

associated with applying these methods separately is frequently far from optimal.

RESULTS: We detail a computational mutagenesis technique based on a four-body,

knowledge-based, statistical contact potential. For any mutation due to a single

amino acid replacement in a protein, the method provides an empirical normalized

measure of the ensuing environmental perturbation occurring at every residue

position. A feature vector is generated for the mutant by considering

perturbations at the mutated position and it's ordered six nearest neighbors in

the 3-dimensional (3D) protein structure. These predictors of stability change

are evaluated by applying machine learning tools to large training sets of

mutants derived from diverse proteins that have been experimentally studied and

described. Predictive models based on our combined approach are either comparable

to, or in many cases significantly outperform, previously published results.

AVAILABILITY: A web server with supporting documentation is available at

http://proteins.gmu.edu/automute.

DOI: 10.1093/bioinformatics/btn353

PMID: 18632749 [Indexed for MEDLINE]

2367. BMC Genomics. 2008 Sep 15;9:414. doi: 10.1186/1471-2164-9-414.

BOV--a web-based BLAST output visualization tool.

Gollapudi R(1), Revanna KV, Hemmerich C, Schaack S, Dong Q.

Author information:

(1)Center for Genomics and Bioinformatics, Indiana University, Bloomington,

Indiana, USA. rajgolla@indiana.edu

BACKGROUND: The BLAST program is one of the most widely used sequence similarity

search tools for genomic research, even by those biologists lacking extensive

bioinformatics training. As the availability of sequence data increases, more

researchers are downloading the BLAST program for local installation and

performing larger and more complex tasks, including batch queries. In order to

manage and interpret the results of batch queries, a host of software packages

have been developed to assist with data management and post-processing. Among

these programs, there is almost a complete lack of visualization tools to provide

graphic representation of complex BLAST pair-wise alignments. We have developed a

web-based program, BLAST Output Visualization Tool (BOV), that allows users to

interactively visualize the matching regions of query and database hit sequences,

thereby allowing the user to quickly and easily dissect complex matching

patterns.

RESULTS: Users can upload the standard BLAST output in pair-wise alignment format

as input to the web server (including batch queries generated installing and

running the stand-alone BLAST program on a local server). The program extracts

the alignment coordinates of matching regions between the query and the

corresponding database hit sequence. The coordinates are used to plot each

matching region as colored lines or trapezoids. Using the straightforward control

panels throughout the web site, each plotted matching region can be easily

explored in detail by, for example, highlighting the region of interest or

examining the raw pair-wise sequence alignment. Tutorials are provided at the

website to guide users step-by-step through the functional features of BOV.

CONCLUSION: BOV provides a user-friendly web interface to visualize the standard

BLAST output for investigating wide-ranging genomic problems, including single

query and batch query datasets. In particular, this software is valuable to users

interested in identifying regions of co-linearity, duplication, translocation,

and inversion among sequences. A web server hosting BOV is accessible via

http://bioportal.cgb.indiana.edu/cgi-bin/BOV/index.cgi and the software is freely

available for local installations.

DOI: 10.1186/1471-2164-9-414

PMCID: PMC2566317

PMID: 18793422 [Indexed for MEDLINE]

2368. Bioinformatics. 2008 Sep 1;24(17):1953-4. doi: 10.1093/bioinformatics/btn341.

Epub 2008 Jul 19.

PIGS: automatic prediction of antibody structures.

Marcatili P(1), Rosi A, Tramontano A.

Author information:

(1)Department of Biochemical Sciences and Istituto Pasteur Fondazione Cenci

Bolognetti, Sapienza University, P.le A. Moro 5, 00185 Rome, Italy.

We describe a web server for the automatic prediction of immunoglobulin variable

domains based on the canonical structure model. The server is user-friendly and

flexible. It allows the user to select the templates for the frameworks and the

loops using different strategies. The final output is a full-fledged 3D model of

the variable domains of the target immunoglobulin.AVAILABILITY: The server is

openly accessible to academic users at the address:

http://arianna.bio.uniroma1.it/pigs. It does not require registration and there

is no limit to the number of sequences that can be submitted.

DOI: 10.1093/bioinformatics/btn341

PMID: 18641403 [Indexed for MEDLINE]

2369. Hear Res. 2008 Sep;243(1-2):11-7. doi: 10.1016/j.heares.2008.04.014. Epub 2008

May 25.

Development of the mouse cochlea database (MCD).

Santi PA(1), Rapson I, Voie A.

Author information:

(1)Department of Otolaryngology, University of Minnesota, Minneapolis, MN 55455,

USA. psanti@umn.edu

The mouse cochlea database (MCD) provides an interactive, image database of the

mouse cochlea for learning its anatomy and data mining of its resources. The MCD

website is hosted on a centrally maintained, high-speed server at the following

URL: (http://mousecochlea.umn.edu). The MCD contains two types of image

resources, serial 2D image stacks and 3D reconstructions of cochlear structures.

Complete image stacks of the cochlea from two different mouse strains were

obtained using orthogonal plane fluorescence optical microscopy (OPFOS). 2D

images of the cochlea are presented on the MCD website as: viewable images within

a stack, 2D atlas of the cochlea, orthogonal sections, and direct volume

renderings combined with isosurface reconstructions. In order to assess cochlear

structures quantitatively, "true" cross-sections of the scala media along the

length of the basilar membrane were generated by virtual resectioning of a

cochlea orthogonal to a cochlear structure, such as the centroid of the basilar

membrane or the scala media. 3D images are presented on the MCD website as:

direct volume renderings, movies, interactive QuickTime VRs, flythrough, and

isosurface 3D reconstructions of different cochlear structures. 3D computer

models can also be used for solid model fabrication by rapid prototyping and

models from different cochleas can be combined to produce an average 3D model.

The MCD is the first comprehensive image resource on the mouse cochlea and is a

new paradigm for understanding the anatomy of the cochlea, and establishing

morphometric parameters of cochlear structures in normal and mutant mice.

DOI: 10.1016/j.heares.2008.04.014

PMCID: PMC2628570

PMID: 18603386 [Indexed for MEDLINE]

2370. IET Syst Biol. 2008 Sep;2(5):352-62. doi: 10.1049/iet-syb:20080102.

Virtual Cell modelling and simulation software environment.

Moraru II(1), Schaff JC, Slepchenko BM, Blinov ML, Morgan F, Lakshminarayana A,

Gao F, Li Y, Loew LM.

Author information:

(1)University of Connecticut Health Center, Center of Cell Analysis and Modeling,

Connecticut, CA 06030, USA.

The Virtual Cell (VCell; http://vcell.org/) is a problem solving environment,

built on a central database, for analysis, modelling and simulation of cell

biological processes. VCell integrates a growing range of molecular mechanisms,

including reaction kinetics, diffusion, flow, membrane transport, lateral

membrane diffusion and electrophysiology, and can associate these with geometries

derived from experimental microscope images. It has been developed and deployed

as a web-based, distributed, client-server system, with more than a thousand

world-wide users. VCell provides a separation of layers (core technologies and

abstractions) representing biological models, physical mechanisms, geometry,

mathematical models and numerical methods. This separation clarifies the impact

of modelling decisions, assumptions and approximations. The result is a

physically consistent, mathematically rigorous, spatial modelling and simulation

framework. Users create biological models and VCell will automatically (i)

generate the appropriate mathematical encoding for running a simulation and (ii)

generate and compile the appropriate computer code. Both deterministic and

stochastic algorithms are supported for describing and running non-spatial

simulations; a full partial differential equation solver using the finite volume

numerical algorithm is available for reaction-diffusion-advection simulations in

complex cell geometries including 3D geometries derived from microscope images.

Using the VCell database, models and model components can be reused and updated,

as well as privately shared among collaborating groups, or published. Exchange of

models with other tools is possible via import/export of SBML, CellML and MatLab

formats. Furthermore, curation of models is facilitated by external database

binding mechanisms for unique identification of components and by standardised

annotations compliant with the MIRIAM standard. VCell is now open source, with

its native model encoding language (VCML) being a public specification, which

stands as the basis for a new generation of more customised, experiment-centric

modelling tools using a new plug-in based platform.

DOI: 10.1049/iet-syb:20080102

PMCID: PMC2711391

PMID: 19045830 [Indexed for MEDLINE]

2371. Int J Eat Disord. 2008 Sep;41(6):527-34. doi: 10.1002/eat.20542.

Online discussion groups for bulimia nervosa: an inductive approach to

Internet-based communication between patients.

Wesemann D(1), Grunwald M.

Author information:

(1)Faculty of Medicine, Institut for Brain Research, University of Leipzig,

Leipzig, Germany. dorette.wesemann@medizin.uni-leipzig.de

OBJECTIVE: Online discussion forums are often used by people with eating

disorders.

METHOD: This study analyses 2,072 threads containing a total of 14,903 postings

from an unmoderated German "prorecovery" forum for persons suffering from bulimia

nervosa (www.ab-server.de) during the period from October 2004 to May 2006. The

threads were inductively analyzed for underlying structural types, and the

various types found were then analyzed for differences in temporal and

quantitative parameters.

RESULTS: Communication in the online discussion forum occurred in three types of

thread: (1) problem-oriented threads (78.8% of threads), (2)

communication-oriented threads (15.3% of threads), and (3) metacommunication

threads (2.6% of threads). Metacommunication threads contained significantly more

postings than problem-oriented and communication-oriented threads, and they were

viewed significantly more often. Moreover, there are temporal differences between

the structural types.

CONCLUSION: Topics relating to active management of the disorder receive great

attention in prorecovery forums.

(c) 2008 by Wiley Periodicals, Inc.

DOI: 10.1002/eat.20542

PMID: 18433031 [Indexed for MEDLINE]

2372. Int J Legal Med. 2008 Sep;122(5):435-40. doi: 10.1007/s00414-008-0233-7. Epub

2008 May 20.

The SNPforID browser: an online tool for query and display of frequency data from

the SNPforID project.

Amigo J(1), Phillips C, Lareu M, Carracedo A.

Author information:

(1)Spanish National Genotyping Center (CeGen) and Genomic Medicine Group,

CIBERER, University of Santiago de Compostela, Santiago de Compostela, Spain.

jamigo@usc.es

The SNPforID browser is a web-based tool for the query and visualization of the

SNP allele frequency data generated by the SNPforID consortium (

http://www.snpforid.org/ ). From this project, validated panels of single

nucleotide polymorphisms (SNPs) for a variety of forensic applications have been

generated with the browser concentrating on the single-tube identification SNP

set comprising 52 markers. A web interface allows the visitor to review the

allele frequencies of the studied markers from all the available populations used

by SNPforID to validate global SNP variability. The interface has been designed

to offer the useful facility of combining populations into appropriate geographic

groups for visual comparison of populations individually or amongst user-defined

groupings and with equivalent HapMap data.

DOI: 10.1007/s00414-008-0233-7

PMID: 18491122 [Indexed for MEDLINE]

2373. J Chem Inf Model. 2008 Sep;48(9):1903-8. doi: 10.1021/ci800178a. Epub 2008 Sep 3.

LOCUSTRA: accurate prediction of local protein structure using a two-layer

support vector machine approach.

Zimmermann O(1), Hansmann UH.

Author information:

(1)John von Neumann Institut für Computing, Research Centre Jülich, 52425 Jülich,

Germany. olav.zimmermann@fz-juelich.de

Constraint generation for 3d structure prediction and structure-based database

searches benefit from fine-grained prediction of local structure. In this work,

we present LOCUSTRA, a novel scheme for the multiclass prediction of local

structure that uses two layers of support vector machines (SVM). Using a

16-letter structural alphabet from de Brevern et al. (Proteins: Struct., Funct.,

Bioinf. 2000, 41, 271-287), we assess its prediction ability for an independent

test set of 222 proteins and compare our method to three-class secondary

structure prediction and direct prediction of dihedral angles. The prediction

accuracy is Q16=61.0% for the 16 classes of the structural alphabet and Q3=79.2%

for a simple mapping to the three secondary classes helix, sheet, and coil. We

achieve a mean phi(psi) error of 24.74 degrees (38.35 degrees) and a median RMSDA

(root-mean-square deviation of the (dihedral) angles) per protein chain of 52.1

degrees. These results compare favorably with related approaches. The LOCUSTRA

web server is freely available to researchers at

http://www.fz-juelich.de/nic/cbb/service/service.php.

DOI: 10.1021/ci800178a

PMID: 18763837 [Indexed for MEDLINE]

2374. OMICS. 2008 Sep;12(3):217-26. doi: 10.1089/omi.2008.0026.

SNP ID-info: SNP ID searching and visualization platform.

Yang CH(1), Chuang LY, Cheng YH, Wen CH, Chang PL, Chang HW.

Author information:

(1)Department of Electronic Engineering, National Kaohsiung University of Applied

Sciences, Taiwan, ROC.

Many association studies provide the relationship between single nucleotide

polymorphisms (SNPs), diseases and cancers, without giving a SNP ID, however.

Here, we developed the SNP ID-info freeware to provide the SNP IDs within

inputting genetic and physical information of genomes. The program provides an

"SNP-ePCR" function to generate the full-sequence using primers and template

inputs. In "SNPosition," sequence from SNP-ePCR or direct input is fed to match

the SNP IDs from SNP fasta-sequence. In "SNP search" and "SNP fasta" function,

information of SNPs within the cytogenetic band, contig position, and keyword

input are acceptable. Finally, the SNP ID neighboring environment for inputs is

completely visualized in the order of contig position and marked with SNP and

flanking hits. The SNP identification problems inherent in NCBI SNP BLAST are

also avoided. In conclusion, the SNP ID-info provides a visualized SNP ID

environment for multiple inputs and assists systematic SNP association studies.

The server and user manual are available at http://bio.kuas.edu.tw/snpid-info.

DOI: 10.1089/omi.2008.0026

PMID: 18582176 [Indexed for MEDLINE]

2375. Protein Sci. 2008 Sep;17(9):1505-12. doi: 10.1110/ps.035691.108. Epub 2008 Jun 2.

Position-specific residue preference features around the ends of helices and

strands and a novel strategy for the prediction of secondary structures.

Duan M(1), Huang M, Ma C, Li L, Zhou Y.

Author information:

(1)Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University

of Science and Technology,Wuhan 430074, People's Republic of China.

It has been many years since position-specific residue preference around the ends

of a helix was revealed. However, all the existing secondary structure prediction

methods did not exploit this preference feature, resulting in low accuracy in

predicting the ends of secondary structures. In this study, we collected a

relatively large data set consisting of 1860 high-resolution, non-homology

proteins from the PDB, and further analyzed the residue distributions around the

ends of regular secondary structures. It was found that there exist

position-specific residue preferences (PSRP) around the ends of not only helices

but also strands. Based on the unique features, we proposed a novel strategy and

developed a tool named E-SSpred that treats the secondary structure as a whole

and builds models to predict entire secondary structure segments directly by

integrating relevant features. In E-SSpred, the support vector machine (SVM)

method is adopted to model and predict the ends of helices and strands according

to the unique residue distributions around them. A simple linear discriminate

analysis method is applied to model and predict entire secondary structure

segments by integrating end-prediction results, tri-peptide composition, and

length distribution features of secondary structures, as well as the prediction

results of the most famous program PSIPRED. The results of fivefold

cross-validation on a widely used data set demonstrate that the accuracy of

E-SSpred in predicting ends of secondary structures is about 10% higher than

PSIPRED, and the overall prediction accuracy (Q(3) value) of E-SSpred (82.2%) is

also better than PSIPRED (80.3%). The E-SSpred web server is available at

http://bioinfo.hust.edu.cn/bio/tools/E-SSpred/index.html.

DOI: 10.1110/ps.035691.108

PMCID: PMC2525534

PMID: 18519808 [Indexed for MEDLINE]

2376. Radiographics. 2008 Sep-Oct;28(5):1251-8. doi: 10.1148/rg.285085701. Epub 2008

Jul 6.

An inexpensive distance learning solution for delivering high-quality live

broadcasts.

Jeun BS(1), Javan R, Gay SB, Olazagasti JM, Bassignani MJ.

Author information:

(1)Department of Radiology, University of Virginia Health System, 1215 Lee St,

Charlottesville, VA 22903, USA. bjeun@virginia.edu

Providing an adequate method of distance learning is a challenge faced by many

multicenter residency programs. The delivery of live didactics over the Internet

is a convenient means of providing a uniform and equivalent educational

experience to residents at distant sites. An application called MedCast has been

developed with use of existing technologies, without the need for costly

commercial products or equipment. MedCast captures the presenter's computer

screen and audio from a microphone source to produce a streaming video that is

transmitted online and archived on a local server. Offsite residents can view

broadcasts in real time or access archived conference sessions for later viewing.

MedCast is available for download at no cost and offers several advantages,

including a user-friendly graphical display interface, near-perfect preservation

of image quality, and cost efficiency. Future plans include objective assessment

of the efficacy of MedCast by comparing postlecture examinations to help evaluate

for any differences between on- and offsite residents in terms of knowledge

gained. A movie clip to supplement this article is available online at

http://radiographics.rsnajnls.org/cgi/content/full/285085701/DC1.

(c) RSNA, 2008.

DOI: 10.1148/rg.285085701

PMID: 18603661 [Indexed for MEDLINE]

2377. J Integr Bioinform. 2008 Aug 25;5(2). doi: 10.2390/biecoll-jib-2008-104.

GOblet: annotation of anonymous sequence data with gene ontology and pathway

terms.

Groth D(1), Hartmann S, Panopoulou G, Poustka AJ, Hennig S.

Author information:

(1)Potsdam University, Bioinformatics Group, c/o MPIMP, Am Mühlenberg 1, D-14476

Potsdam-Golm, Germany. dgroth@mpimp-golm.mpg.de

The functional annotation of genomic data has become a major task for the

ever-growing number of sequencing projects. In order to address this challenge,

we recently developed GOblet, a free web service for the annotation of anonymous

sequences with Gene Ontology (GO) terms. However, to overcome limitations of the

GO terminology, and to aid in understanding not only single components but as

well systemic interactions between the individual components, we have now

extended the GOblet web service to integrate also pathway annotations.

Furthermore, we extended and upgraded the data analysis pipeline with improved

summaries, and added term enrichment and clustering algorithms. Finally, we are

now making GOblet available as a stand-alone application for high-throughput

processing on local machines. The advantages of this frequently requested feature

is that a) the user can avoid restrictions of our web service for uploading and

processing large amounts of data, and that b) confidential data can be analysed

without insecure transfer to a public web server. The stand-alone version of the

web service has been implemented using platform independent Tcl-scripts, which

can be run with just a single runtime file utilizing the Starkit technology. The

GOblet web service and the stand-alone application are freely available at

http://goblet.molgen.mpg.de.

DOI: 10.2390/biecoll-jib-2008-104

PMID: 20134064 [Indexed for MEDLINE]

2378. J Integr Bioinform. 2008 Aug 25;5(2). doi: 10.2390/biecoll-jib-2008-91.

MoRAine--a web server for fast computational transcription factor binding motif

re-annotation.

Baumbach J(1), Wittkop T, Weile J, Kohl T, Rahmann S.

Author information:

(1)Genome Informatics, Faculty of Technology, Bielefeld University, D-33594

Bielefeld, Germany. Jan.Baumbach@CeBiTec.Uni-Bielefeld.DE

BACKGROUND: A precise experimental identification of transcription factor binding

motifs (TFBMs), accurate to a single base pair, is time-consuming and diffcult.

For several databases, TFBM annotations are extracted from the literature and

stored 5' --> 3' relative to the target gene. Mixing the two possible

orientations of a motif results in poor information content of subsequently

computed position frequency matrices (PFMs) and sequence logos. Since these PFMs

are used to predict further TFBMs, we address the question if the TFBMs

underlying a PFM can be re-annotated automatically to improve both the

information content of the PFM and subsequent classification performance.

RESULTS: We present MoRAine, an algorithm that re-annotates transcription factor

binding motifs. Each motif with experimental evidence underlying a PFM is

compared against each other such motif. The goal is to re-annotate TFBMs by

possibly switching their strands and shifting them a few positions in order to

maximize the information content of the resulting adjusted PFM. We present two

heuristic strategies to perform this optimization and subsequently show that

MoRAine significantly improves the corresponding sequence logos. Furthermore, we

justify the method by evaluating specificity, sensitivity, true positive, and

false positive rates of PFM-based TFBM predictions for E. coli using the original

database motifs and the MoRAine-adjusted motifs. The classification performance

is considerably increased if MoRAine is used as a preprocessing step.

CONCLUSIONS: MoRAine is integrated into a publicly available web server and can

be used online or downloaded as a stand-alone version from

http://moraine.cebitec. uni-bielefeld.de.

DOI: 10.2390/biecoll-jib-2008-91

PMID: 20134062 [Indexed for MEDLINE]

2379. BMC Res Notes. 2008 Aug 21;1:67. doi: 10.1186/1756-0500-1-67.

GPCRTree: online hierarchical classification of GPCR function.

Davies MN(1), Secker A, Halling-Brown M, Moss DS, Freitas AA, Timmis J, Clark E,

Flower DR.

Author information:

(1)The Jenner Institute, University of Oxford, Compton, Newbury, Berkshire, RG20

7NN, UK. m.davies@mail.cryst.bbk.ac.uk

BACKGROUND: G protein-coupled receptors (GPCRs) play important physiological

roles transducing extracellular signals into intracellular responses.

Approximately 50% of all marketed drugs target a GPCR. There remains considerable

interest in effectively predicting the function of a GPCR from its primary

sequence.

FINDINGS: Using techniques drawn from data mining and proteochemometrics, an

alignment-free approach to GPCR classification has been devised. It uses a simple

representation of a protein's physical properties. GPCRTree, a publicly-available

internet server, implements an algorithm that classifies GPCRs at the class,

sub-family and sub-subfamily level.

CONCLUSION: A selective top-down classifier was developed which assigns sequences

within a GPCR hierarchy. Compared to other publicly available GPCR prediction

servers, GPCRTree is considerably more accurate at every level of classification.

The server has been available online since March 2008 at URL:

http://igrid-ext.cryst.bbk.ac.uk/gpcrtree/.

DOI: 10.1186/1756-0500-1-67

PMCID: PMC2547103

PMID: 18717986

2380. BMC Bioinformatics. 2008 Aug 18;9:344. doi: 10.1186/1471-2105-9-344.

An enhanced partial order curve comparison algorithm and its application to

analyzing protein folding trajectories.

Sun H(1), Ferhatosmanoglu H, Ota M, Wang Y.

Author information:

(1)Department of Computer Science and Engineering, The Ohio State University,

Columbus, OH 43210, USA. sun.82@osu.edu

BACKGROUND: Understanding how proteins fold is essential to our quest in

discovering how life works at the molecular level. Current computation power

enables researchers to produce a huge amount of folding simulation data. Hence

there is a pressing need to be able to interpret and identify novel folding

features from them.

RESULTS: In this paper, we model each folding trajectory as a multi-dimensional

curve. We then develop an effective multiple curve comparison (MCC) algorithm,

called the enhanced partial order (EPO) algorithm, to extract features from a set

of diverse folding trajectories, including both successful and unsuccessful

simulation runs. The EPO algorithm addresses several new challenges presented by

comparing high dimensional curves coming from folding trajectories. A detailed

case study on miniprotein Trp-cage 1 demonstrates that our algorithm can detect

similarities at rather low level, and extract biologically meaningful folding

events.

CONCLUSION: The EPO algorithm is general and applicable to a wide range of

applications. We demonstrate its generality and effectiveness by applying it to

aligning multiple protein structures with low similarities. For user's

convenience, we provide a web server for the algorithm at

http://db.cse.ohio-state.edu/EPO.

DOI: 10.1186/1471-2105-9-344

PMCID: PMC2571979

PMID: 18710565 [Indexed for MEDLINE]

2381. Bioinformatics. 2008 Aug 15;24(16):i112-8. doi: 10.1093/bioinformatics/btn288.

RNA structure alignment by a unit-vector approach.

Capriotti E(1), Marti-Renom MA.

Author information:

(1)Bioinformatics and Genomics Department, Structural Genomics Unit, Centro de

Investigación Príncipe Felipe, Valencia, Spain.

MOTIVATION: The recent discovery of tiny RNA molecules such as microRNAs and

small interfering RNA are transforming the view of RNA as a simple information

transfer molecule. Similar to proteins, the native three-dimensional structure of

RNA determines its biological activity. Therefore, classifying the current

structural space is paramount for functionally annotating RNA molecules. The

increasing numbers of RNA structures deposited in the PDB requires more accurate,

automatic and benchmarked methods for RNA structure comparison. In this article,

we introduce a new algorithm for RNA structure alignment based on a unit-vector

approach. The algorithm has been implemented in the SARA program, which results

in RNA structure pairwise alignments and their statistical significance.

RESULTS: The SARA program has been implemented to be of general applicability

even when no secondary structure can be calculated from the RNA structures. A

benchmark against the ARTS program using a set of 1275 non-redundant pairwise

structure alignments results in inverted approximately 6% extra alignments with

at least 50% structurally superposed nucleotides and base pairs. A first attempt

to perform RNA automatic functional annotation based on structure alignments

indicates that SARA can correctly assign the deepest SCOR classification to >60%

of the query structures.

AVAILABILITY: The SARA program is freely available through a World Wide Web

server http://sgu.bioinfo.cipf.es/services/SARA/.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btn288

PMID: 18689811 [Indexed for MEDLINE]

2382. Bioinformatics. 2008 Aug 15;24(16):1819-20. doi: 10.1093/bioinformatics/btn255.

Epub 2008 Jun 10.

MINS2: revisiting the molecular code for transmembrane-helix recognition by the

Sec61 translocon.

Park Y(1), Helms V.

Author information:

(1)Center for Bioinformatics, Saarland University, Germany.

To be fully functional, membrane proteins should not only fold, but also get

inserted into the membrane, which is mediated by the Sec61 translocon. Recent

experimental studies have attempted to elucidate how the Sec61 translocon

accomplishes this delicate task by measuring the translocon-mediated membrane

insertion free energies of 357 systematically designed peptides. On the basis of

this data set, we have developed MINS2, a novel sequence-based computational

method for predicting the membrane insertion free energies of protein sequences.

A benchmark analysis of MINS2 shows that MINS2 signi.cantly outperforms

previously proposed methods. Importantly, the application of MINS2 to known

membrane protein structures shows that a better prediction of membrane insertion

free energies does not lead to a better prediction of transmembrane segments of

polytopic membrane proteins.AVAILABILITY: A web server for MINS2 is publicly

available at http://service.bioinformatik.uni-saarland.de/mins.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btn255

PMID: 18544549 [Indexed for MEDLINE]

2383. Biochem Biophys Res Commun. 2008 Aug 8;372(4):831-4. doi:

10.1016/j.bbrc.2008.05.134. Epub 2008 Jun 2.

CID-miRNA: a web server for prediction of novel miRNA precursors in human genome.

Tyagi S(1), Vaz C, Gupta V, Bhatia R, Maheshwari S, Srinivasan A, Bhattacharya A.

Author information:

(1)Centre for Computational Biology & Bioinformatics, School of Information

Technology, Jawaharlal Nehru University, New Mehrauli Road, New Delhi 110067,

India.

microRNAs (miRNA) are a class of non-protein coding functional RNAs that are

thought to regulate expression of target genes by direct interaction with mRNAs.

miRNAs have been identified through both experimental and computational methods

in a variety of eukaryotic organisms. Though these approaches have been partially

successful, there is a need to develop more tools for detection of these RNAs as

they are also thought to be present in abundance in many genomes. In this report

we describe a tool and a web server, named CID-miRNA, for identification of miRNA

precursors in a given DNA sequence, utilising secondary structure-based filtering

systems and an algorithm based on stochastic context free grammar trained on

human miRNAs. CID-miRNA analyses a given sequence using a web interface, for

presence of putative miRNA precursors and the generated output lists all the

potential regions that can form miRNA-like structures. It can also scan large

genomic sequences for the presence of potential miRNA precursors in its

stand-alone form. The web server can be accessed at

http://mirna.jnu.ac.in/cidmirna/.

DOI: 10.1016/j.bbrc.2008.05.134

PMID: 18522801 [Indexed for MEDLINE]

2384. BMC Bioinformatics. 2008 Aug 7;9:333. doi: 10.1186/1471-2105-9-333.

The SeqWord Genome Browser: an online tool for the identification and

visualization of atypical regions of bacterial genomes through oligonucleotide

usage.

Ganesan H(1), Rakitianskaia AS, Davenport CF, Tümmler B, Reva ON.

Author information:

(1)Dep of Biochemistry, Bioinformatics and Computational Biology Unit, University

of Pretoria, Lynnwood road, Hillcrest, Pretoria, 0002, South Africa.

ham.kzn@gmail.com

BACKGROUND: Data mining in large DNA sequences is a major challenge in microbial

genomics and bioinformatics. Oligonucleotide usage (OU) patterns provide a wealth

of information for large scale sequence analysis and visualization. The purpose

of this research was to make OU statistical analysis available as a novel

web-based tool for functional genomics and annotation. The tool is also available

as a downloadable package.

RESULTS: The SeqWord Genome Browser (SWGB) was developed to visualize the natural

compositional variation of DNA sequences. The applet is also used for

identification of divergent genomic regions both in annotated sequences of

bacterial chromosomes, plasmids, phages and viruses, and in raw DNA sequences

prior to annotation by comparing local and global OU patterns. The applet allows

fast and reliable identification of clusters of horizontally transferred genomic

islands, large multi-domain genes and genes for ribosomal RNA. Within the

majority of genomic fragments (also termed genomic core sequence), regions

enriched with housekeeping genes, ribosomal proteins and the regions rich in

pseudogenes or genetic vestiges may be contrasted.

CONCLUSION: The SWGB applet presents a range of comprehensive OU statistical

parameters calculated for a range of bacterial species, plasmids and phages. It

is available on the Internet at http://www.bi.up.ac.za/SeqWord/mhhapplet.php.

DOI: 10.1186/1471-2105-9-333

PMCID: PMC2528017

PMID: 18687122 [Indexed for MEDLINE]

2385. Amino Acids. 2008 Aug;35(2):295-302. doi: 10.1007/s00726-007-0634-9. Epub 2008

Jan 31.

PRINTR: prediction of RNA binding sites in proteins using SVM and profiles.

Wang Y(1), Xue Z, Shen G, Xu J.

Author information:

(1)Institute of Biophysics and Biochemistry, School of Life Science, Huazhong

University of Science and Technology, Wuhan City, China. yanw@mail.hust.edu.cn

Protein-RNA interactions play a key role in a number of biological processes such

as protein synthesis, mRNA processing, assembly and function of ribosomes and

eukaryotic spliceosomes. A reliable identification of RNA-binding sites in

RNA-binding proteins is important for functional annotation and site-directed

mutagenesis. We developed a novel method for the prediction of protein residues

that interact with RNA using support vector machine (SVM) and position-specific

scoring matrices (PSSMs). Two cases have been considered in the prediction of

protein residues at RNA-binding surfaces. One is given the sequence information

of a protein chain that is known to interact with RNA; the other is given the

structural information. Thus, five different inputs have been tested. Coupled

with PSI-BLAST profiles and predicted secondary structure, the present approach

yields a Matthews correlation coefficient (MCC) of 0.432 by a 7-fold

cross-validation, which is the best among all previous reported RNA-binding sites

prediction methods. When given the structural information, we have obtained the

MCC value of 0.457, with PSSMs, observed secondary structure and solvent

accessibility information assigned by DSSP as input. A web server implementing

the prediction method is available at the following URL:

http://210.42.106.80/printr/ .

DOI: 10.1007/s00726-007-0634-9

PMID: 18235992 [Indexed for MEDLINE]

2386. Amino Acids. 2008 Aug;35(2):345-53. Epub 2007 Dec 28.

AAIndexLoc: predicting subcellular localization of proteins based on a new

representation of sequences using amino acid indices.

Tantoso E(1), Li KB.

Author information:

(1)Bioinformatics Institute, Singapore.

Identifying a protein's subcellular localization is an important step to

understand its function. However, the involved experimental work is usually

laborious, time consuming and costly. Computational prediction hence becomes

valuable to reduce the inefficiency. Here we provide a method to predict protein

subcellular localization by using amino acid composition and physicochemical

properties. The method concatenates the information extracted from a protein's

N-terminal, middle and full sequence. Each part is represented by amino acid

composition, weighted amino acid composition, five-level grouping composition and

five-level dipeptide composition. We divided our dataset into training and

testing set. The training set is used to determine the best performing amino acid

index by using five-fold cross validation, whereas the testing set acts as the

independent dataset to evaluate the performance of our model. With the novel

representation method, we achieve an accuracy of approximately 75% on independent

dataset. We conclude that this new representation indeed performs well and is

able to extract the protein sequence information. We have developed a web server

for predicting protein subcellular localization. The web server is available at

http://aaindexloc.bii.a-star.edu.sg .

DOI: 10.1007/s00726-007-0616-y

PMID: 18163182 [Indexed for MEDLINE]

2387. Bioinformatics. 2008 Aug 1;24(15):1731-2. doi: 10.1093/bioinformatics/btn259.

Epub 2008 Jun 9.

DNAlive: a tool for the physical analysis of DNA at the genomic scale.

Goñi JR(1), Fenollosa C, Pérez A, Torrents D, Orozco M.

Author information:

(1)Joint IRB-BSC Program on Computational Biology, Institute of Research in

Biomedicine, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028,

Spain.

SUMMARY: DNAlive is a tool for the analysis and graphical display of structural

and physical characteristics of genomic DNA. The web server implements a wide

repertoire of metrics to derive physical information from DNA sequences with a

powerful interface to derive 3D information on large sequences of both naked and

protein-bound DNAs. Furthermore, it implements a mesoscopic Metropolis code which

allows the inexpensive study of the dynamic properties of chromatin fibers. In

addition, our server also surveys other protein and genomic databases allowing

the user to combine and explore the physical properties of selected DNA in the

context of functional features annotated on those regions.

AVAILABILITY: http://mmb.pcb.ub.es/DNAlive/ ; http://www.inab.org/

DOI: 10.1093/bioinformatics/btn259

PMID: 18544548 [Indexed for MEDLINE]

2388. Bioinformatics. 2008 Aug 1;24(15):1662-8. doi: 10.1093/bioinformatics/btn221.

Epub 2008 May 12.

OCTOPUS: improving topology prediction by two-track ANN-based preference scores

and an extended topological grammar.

Viklund H(1), Elofsson A.

Author information:

(1)Department of Biochemistry and Biophysics/Center for Biomembrane

Research/Stockholm Bioinformatics Center, The Arrhenius Laboratories for Natural

Sciences, Stockholm University, SE-10691 Stockholm, Sweden.

MOTIVATION: As alpha-helical transmembrane proteins constitute roughly 25% of a

typical genome and are vital parts of many essential biological processes,

structural knowledge of these proteins is necessary for increasing our

understanding of such processes. Because structural knowledge of transmembrane

proteins is difficult to attain experimentally, improved methods for prediction

of structural features of these proteins are important.

RESULTS: OCTOPUS, a new method for predicting transmembrane protein topology is

presented and benchmarked using a dataset of 124 sequences with known structures.

Using a novel combination of hidden Markov models and artificial neural networks,

OCTOPUS predicts the correct topology for 94% of the sequences. In particular,

OCTOPUS is the first topology predictor to fully integrate modeling of

reentrant/membrane-dipping regions and transmembrane hairpins in the topological

grammar.

AVAILABILITY: OCTOPUS is available as a web server at http://octopus.cbr.su.se.

DOI: 10.1093/bioinformatics/btn221

PMID: 18474507 [Indexed for MEDLINE]

2389. Comput Biol Chem. 2008 Aug;32(4):298-301. doi:

10.1016/j.compbiolchem.2008.03.010. Epub 2008 Apr 1.

A method for discovering transmembrane beta-barrel proteins in Gram-negative

bacterial proteomes.

Hu J(1), Yan C.

Author information:

(1)Department of Computer Science, Utah State University, Logan, UT 84322, USA.

Transmembrane beta-barrel (TMB) proteins play pivotal roles in many aspects of

bacterial functions. This paper presents a k-nearest neighbor (K-NN) method for

discriminating TMB and non-TMB proteins. We start with a method that makes

predictions based on a distance computed from residue composition and gradually

improve the prediction performance by including homologous sequences and

searching for a set of residues and di-peptides for calculating the distance. The

final method achieves an accuracy of 97.1%, with 0.876 MCC, 86.4% sensitivity and

98.8% specificity. A web server based on the proposed method is available at

http://yanbioinformatics.cs.usu.edu:8080/TMBKNNsubmit.

DOI: 10.1016/j.compbiolchem.2008.03.010

PMID: 18467177 [Indexed for MEDLINE]

2390. Mol Divers. 2008 Aug-Nov;12(3-4):171-9. doi: 10.1007/s11030-008-9093-9. Epub 2008

Oct 25.

Prediction of protein structural classes using hybrid properties.

Li W(1), Lin K, Feng K, Cai Y.

Author information:

(1)CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for

Biological Sciences, Chinese Academy of Sciences, Shanghai, China.

In this paper, amino acid compositions are combined with some protein sequence

properties (physiochemical properties) to predict protein structural classes. We

are able to predict protein structural classes using a mathematical model that

combines the nearest neighbor algorithm (NNA), mRMR (minimum redundancy, maximum

relevance), and feature forward searching strategy. Jackknife cross-validation is

used to evaluate the prediction accuracy. As a result, the prediction success

rate improves to 68.8%, which is better than the 62.2% obtained when using only

amino acid compositions. Therefore, we conclude that the physiochemical

properties are factors that contribute to the protein folding phenomena and the

most contributing features are found to be the amino acid composition. We expect

that prediction accuracy will improve further as more sequence information comes

to light. A web server for predicting the protein structural classes is available

at http://app3.biosino.org:8080/liwenjin/index.jsp.

DOI: 10.1007/s11030-008-9093-9

PMID: 18953662 [Indexed for MEDLINE]

2391. Protein Sci. 2008 Aug;17(8):1374-82. doi: 10.1110/ps.035469.108. Epub 2008 Jun

26.

Novel protein folds and their nonsequential structural analogs.

Guerler A(1), Knapp EW.

Author information:

(1)Department of Chemistry and Biochemistry, Freie Universität Berlin, 14195

Berlin, Germany.

Newly determined protein structures are classified to belong to a new fold, if

the structures are sufficiently dissimilar from all other so far known protein

structures. To analyze structural similarities of proteins, structure alignment

tools are used. We demonstrate that the usage of nonsequential structure

alignment tools, which neglect the polypeptide chain connectivity, can yield

structure alignments with significant similarities between proteins of known

three-dimensional structure and newly determined protein structures that possess

a new fold. The recently introduced protein structure alignment tool, GANGSTA, is

specialized to perform nonsequential alignments with proper assignment of the

secondary structure types by focusing on helices and strands only. In the new

version, GANGSTA+, the underlying algorithms were completely redesigned, yielding

enhanced quality of structure alignments, offering alignment against a larger

database of protein structures, and being more efficient. We applied DaliLite,

TM-align, and GANGSTA+ on three protein crystal structures considered to be novel

folds. Applying GANGSTA+ to these novel folds, we find proteins in the ASTRAL40

database, which possess significant structural similarities, albeit the

alignments are nonsequential and in some cases involve secondary structure

elements aligned in reverse orientation. A web server is available at

http://agknapp.chemie.fu-berlin.de/gplus for pairwise alignment, visualization,

and database comparison.

DOI: 10.1110/ps.035469.108

PMCID: PMC2492825

PMID: 18583523 [Indexed for MEDLINE]

2392. Proteins. 2008 Aug;72(2):693-710. doi: 10.1002/prot.21944.

PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides

and probabilistic latent semantic analysis.

Chang JM(1), Su EC, Lo A, Chiu HS, Sung TY, Hsu WL.

Author information:

(1)Bioinformatics Lab, Institute of Information Science, Academia Sinica, Taipei,

Taiwan.

Prediction of protein subcellular localization (PSL) is important for genome

annotation, protein function prediction, and drug discovery. Many computational

approaches for PSL prediction based on protein sequences have been proposed in

recent years for Gram-negative bacteria. We present PSLDoc, a method based on

gapped-dipeptides and probabilistic latent semantic analysis (PLSA) to solve this

problem. A protein is considered as a term string composed by gapped-dipeptides,

which are defined as any two residues separated by one or more positions. The

weighting scheme of gapped-dipeptides is calculated according to a position

specific score matrix, which includes sequence evolutionary information. Then,

PLSA is applied for feature reduction, and reduced vectors are input to five

one-versus-rest support vector machine classifiers. The localization site with

the highest probability is assigned as the final prediction. It has been reported

that there is a strong correlation between sequence homology and subcellular

localization (Nair and Rost, Protein Sci 2002;11:2836-2847; Yu et al., Proteins

2006;64:643-651). To properly evaluate the performance of PSLDoc, a target

protein can be classified into low- or high-homology data sets. PSLDoc's overall

accuracy of low- and high-homology data sets reaches 86.84% and 98.21%,

respectively, and it compares favorably with that of CELLO II (Yu et al.,

Proteins 2006;64:643-651). In addition, we set a confidence threshold to achieve

a high precision at specified levels of recall rates. When the confidence

threshold is set at 0.7, PSLDoc achieves 97.89% in precision which is

considerably better than that of PSORTb v.2.0 (Gardy et al., Bioinformatics

2005;21:617-623). Our approach demonstrates that the specific feature

representation for proteins can be successfully applied to the prediction of

protein subcellular localization and improves prediction accuracy. Besides,

because of the generality of the representation, our method can be extended to

eukaryotic proteomes in the future. The web server of PSLDoc is publicly

available at http://bio-cluster.iis.sinica.edu.tw/~ bioapp/PSLDoc/.

(c) 2008 Wiley-Liss, Inc.

DOI: 10.1002/prot.21944

PMID: 18260102 [Indexed for MEDLINE]

2393. Proteins. 2008 Aug;72(2):547-56. doi: 10.1002/prot.21945.

MUSTER: Improving protein sequence profile-profile alignments by using multiple

sources of structure information.

Wu S(1), Zhang Y.

Author information:

(1)Center for Bioinformatics and Department of Molecular Bioscience, University

of Kansas, 2030 Becker Dr, Lawrence, Kansas 66047, USA.

We develop a new threading algorithm MUSTER by extending the previous sequence

profile-profile alignment method, PPA. It combines various sequence and structure

information into single-body terms which can be conveniently used in dynamic

programming search: (1) sequence profiles; (2) secondary structures; (3)

structure fragment profiles; (4) solvent accessibility; (5) dihedral torsion

angles; (6) hydrophobic scoring matrix. The balance of the weighting parameters

is optimized by a grading search based on the average TM-score of 111 training

proteins which shows a better performance than using the conventional

optimization methods based on the PROSUP database. The algorithm is tested on 500

nonhomologous proteins independent of the training sets. After removing the

homologous templates with a sequence identity to the target >30%, in 224 cases,

the first template alignment has the correct topology with a TM-score >0.5. Even

with a more stringent cutoff by removing the templates with a sequence identity

>20% or detectable by PSI-BLAST with an E-value <0.05, MUSTER is able to identify

correct folds in 137 cases with the first model of TM-score >0.5. Dependent on

the homology cutoffs, the average TM-score of the first threading alignments by

MUSTER is 5.1-6.3% higher than that by PPA. This improvement is statistically

significant by the Wilcoxon signed rank test with a P-value < 1.0 x 10(-13),

which demonstrates the effect of additional structural information on the protein

fold recognition. The MUSTER server is freely available to the academic community

at http://zhang.bioinformatics.ku.edu/MUSTER.

(c) 2008 Wiley-Liss, Inc.

DOI: 10.1002/prot.21945

PMCID: PMC2666101

PMID: 18247410 [Indexed for MEDLINE]

2394. J Comput Chem. 2008 Jul 30;29(10):1675-83. doi: 10.1002/jcc.20925.

Analysis and prediction of protein folding rates using quadratic response surface

models.

Huang LT(1), Gromiha MM.

Author information:

(1)Department of Computer Science and Information Engineering, Ming-Dao

University, Changhua 523, Taiwan.

Understanding the relationship between amino acid sequences and folding rates of

proteins is an important task in computational and molecular biology. In this

work, we have systematically analyzed the composition of amino acid residues for

proteins with different ranges of folding rates. We observed that the polar

residues, Asn, Gln, Ser, and Lys, are dominant in fast folding proteins whereas

the hydrophobic residues, Ala, Cys, Gly, and Leu, prefer to be in slow folding

proteins. Further, we have developed a method based on quadratic response surface

models for predicting the folding rates of 77 two- and three-state proteins. Our

method showed a correlation of 0.90 between experimental and predicted protein

folding rates using leave-one-out cross-validation method. The classification of

proteins based on structural class improved the correlation to 0.98 and it is

0.99, 0.98, and 0.96, respectively, for all-alpha, all-beta, and mixed class

proteins. In addition, we have utilized Baysean classification theory for

discriminating two- and three-state proteins, which showed an accuracy of 90%. We

have developed a web server for predicting protein folding rates and it is

available at http://bioinformatics.myweb.hinet.net/foldrate.htm.

(c) 2008 Wiley Periodicals, Inc. J Comput Chem, 2008.

DOI: 10.1002/jcc.20925

PMID: 18351617 [Indexed for MEDLINE]

2395. J Comput Chem. 2008 Jul 30;29(10):1596-604. doi: 10.1002/jcc.20918.

Prediction of protein structural class using novel evolutionary collocation-based

sequence representation.

Chen K(1), Kurgan LA, Ruan J.

Author information:

(1)Department of Electrical and Computer Engineering, ECERF, University of

Alberta, Edmonton, Alberta, Canada.

Knowledge of structural classes is useful in understanding of folding patterns in

proteins. Although existing structural class prediction methods applied virtually

all state-of-the-art classifiers, many of them use a relatively simple protein

sequence representation that often includes amino acid (AA) composition. To this

end, we propose a novel sequence representation that incorporates evolutionary

information encoded using PSI-BLAST profile-based collocation of AA pairs. We

used six benchmark datasets and five representative classifiers to quantify and

compare the quality of the structural class prediction with the proposed

representation. The best, classifier support vector machine achieved 61-96%

accuracy on the six datasets. These predictions were comprehensively compared

with a wide range of recently proposed methods for prediction of structural

classes. Our comprehensive comparison shows superiority of the proposed

representation, which results in error rate reductions that range between 14% and

26% when compared with predictions of the best-performing, previously published

classifiers on the considered datasets. The study also shows that, for the

benchmark dataset that includes sequences characterized by low identity (i.e.,

25%, 30%, and 40%), the prediction accuracies are 20-35% lower than for the other

three datasets that include sequences with a higher degree of similarity. In

conclusion, the proposed representation is shown to substantially improve the

accuracy of the structural class prediction. A web server that implements the

presented prediction method is freely available at

http://biomine.ece.ualberta.ca/Structural\_Class/SCEC.html.

(c) 2008 Wiley Periodicals, Inc. J Comput Chem, 2008.

DOI: 10.1002/jcc.20918

PMID: 18293306 [Indexed for MEDLINE]

2396. J Proteomics. 2008 Jul 21;71(2):245-8. doi: 10.1016/j.jprot.2008.02.005. Epub

2008 Mar 5.

The World-2DPAGE Constellation to promote and publish gel-based proteomics data

through the ExPASy server.

Hoogland C(1), Mostaguir K, Appel RD, Lisacek F.

Author information:

(1)Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, 1

Michel-Servet, CH-1211 Geneva 4, Switzerland. Christine.Hoogland@isb-sib.ch

Since it was launched in 1993, the ExPASy server has been and is still a

reference in the proteomics world. ExPASy users access various databases, many

dedicated tools, and lists of resources, among other services. A significant part

of resources available is devoted to two-dimensional electrophoresis data. Our

latest contribution to the expansion of the pool of on-line proteomics data is

the World-2DPAGE Constellation, accessible at http://world-2dpage.expasy.org/. It

is composed of the established WORLD-2DPAGE List of 2-D PAGE database servers,

the World-2DPAGE Portal that queries simultaneously world-wide proteomics

databases, and the recently created World-2DPAGE Repository. The latter component

is a public standards-compliant repository for gel-based proteomics data linked

to protein identifications published in the literature. It has been set up using

the Make2D-DB package, a software tool that helps building SWISS-2DPAGE-like

databases on one's own Web site. The lack of necessary informatics infrastructure

to build and run a dedicated website is no longer an obstacle to make proteomics

data publicly accessible on the Internet.

DOI: 10.1016/j.jprot.2008.02.005

PMID: 18617148 [Indexed for MEDLINE]

2397. BMC Bioinformatics. 2008 Jul 15;9:310. doi: 10.1186/1471-2105-9-310.

Computational identification of ubiquitylation sites from protein sequences.

Tung CW(1), Ho SY.

Author information:

(1)Institute of Bioinformatics, National Chiao Tung University, Hsinchu 300,

Taiwan. cwtung@livemail.tw

BACKGROUND: Ubiquitylation plays an important role in regulating protein

functions. Recently, experimental methods were developed toward effective

identification of ubiquitylation sites. To efficiently explore more undiscovered

ubiquitylation sites, this study aims to develop an accurate sequence-based

prediction method to identify promising ubiquitylation sites.

RESULTS: We established an ubiquitylation dataset consisting of 157

ubiquitylation sites and 3676 putative non-ubiquitylation sites extracted from

105 proteins in the UbiProt database. This study first evaluates promising

sequence-based features and classifiers for the prediction of ubiquitylation

sites by assessing three kinds of features (amino acid identity, evolutionary

information, and physicochemical property) and three classifiers (support vector

machine, k-nearest neighbor, and NaïveBayes). Results show that the set of used

531 physicochemical properties and support vector machine (SVM) are the best kind

of features and classifier respectively that their combination has a prediction

accuracy of 72.19% using leave-one-out cross-validation.Consequently, an

informative physicochemical property mining algorithm (IPMA) is proposed to

select an informative subset of 531 physicochemical properties. A prediction

system UbiPred was implemented by using an SVM with the feature set of 31

informative physicochemical properties selected by IPMA, which can improve the

accuracy from 72.19% to 84.44%. To further analyze the informative

physicochemical properties, a decision tree method C5.0 was used to acquire

if-then rule-based knowledge of predicting ubiquitylation sites. UbiPred can

screen promising ubiquitylation sites from putative non-ubiquitylation sites

using prediction scores. By applying UbiPred, 23 promising ubiquitylation sites

were identified from an independent dataset of 3424 putative non-ubiquitylation

sites, which were also validated by using the obtained prediction rules.

CONCLUSION: We have proposed an algorithm IPMA for mining informative

physicochemical properties from protein sequences to build an SVM-based

prediction system UbiPred. UbiPred can predict ubiquitylation sites accompanied

with a prediction score each to help biologists in identifying promising sites

for experimental verification. UbiPred has been implemented as a web server and

is available at http://iclab.life.nctu.edu.tw/ubipred.

DOI: 10.1186/1471-2105-9-310

PMCID: PMC2488362

PMID: 18625080 [Indexed for MEDLINE]

2398. Acta Crystallogr A. 2008 Jul;64(Pt 4):465-75. doi: 10.1107/S010876730801341X.

Epub 2008 Jun 17.

Estimated H-atom anisotropic displacement parameters: a comparison between

different methods and with neutron diffraction results.

Munshi P(1), Madsen AØ, Spackman MA, Larsen S, Destro R.

Author information:

(1)University of Western Australia, Australia.

Anisotropic displacement parameters (ADPs) are compared for H atoms estimated

using three recently described procedures, both among themselves and with neutron

diffraction results. The results convincingly demonstrate that all methods are

capable of giving excellent results for several benchmark systems and identify

systematic discrepancies for several atom types. A revised and extended library

of internal H-atom mean-square displacements is presented for use with Madsen's

SHADE web server [J. Appl. Cryst. (2006), 39, 757-758; http://shade.ki.ku.dk],

and the improvement over the original SHADE results is substantial, suggesting

that this is now the most readily and widely applicable of the three approximate

procedures. Using this new library--SHADE2--it is shown that, in line with

expectations, a segmented rigid-body description of the heavy atoms yields only a

small improvement in the agreement with neutron results. The SHADE2 library, now

incorporated in the SHADE web server, is recommended as a routine procedure for

deriving estimates of H-atom ADPs suitable for use in charge-density studies on

molecular crystals, and its widespread use should reveal remaining deficiencies

and perhaps overcome the inherent bias in the majority of such studies.

DOI: 10.1107/S010876730801341X

PMID: 18560163 [Indexed for MEDLINE]

2399. Bioinformatics. 2008 Jul 1;24(13):1536-7. doi: 10.1093/bioinformatics/btn234.

Epub 2008 May 16.

Swelfe: a detector of internal repeats in sequences and structures.

Abraham AL(1), Rocha EP, Pothier J.

Author information:

(1)UPMC Univ Paris 06, Atelier de BioInformatique, F75005 Paris, France.

annela@abi.snv.jussieu.fr

Intragenic duplications of genetic material have important biological roles

because of their protein sequence and structural consequences. We developed

Swelfe to find internal repeats at three levels. Swelfe quickly identifies

statistically significant internal repeats in DNA and amino acid sequences and in

3D structures using dynamic programming. The associated web server also shows the

relationships between repeats at each level and facilitates visualization of the

results.AVAILABILITY: http://bioserv.rpbs.jussieu.fr/swelfe.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btn234

PMCID: PMC2718673

PMID: 18487242 [Indexed for MEDLINE]

2400. Bioinformatics. 2008 Jul 1;24(13):1542-6. doi: 10.1093/bioinformatics/btn203.

Epub 2008 May 14.

An overview of the wcd EST clustering tool.

Hazelhurst S(1), Hide W, Lipták Z, Nogueira R, Starfield R.

Author information:

(1)Wits Bioinformatics, University of the Witwatersrand, Johannesburg, Private

Bag 3, 2050 Wits, South Africa. scott.hazelhurst@wits.ac.za

The wcd system is an open source tool for clustering expressed sequence tags

(EST) and other DNA and RNA sequences. wcd allows efficient all-versus-all

comparison of ESTs using either the d(2) distance function or edit distance,

improving existing implementations of d(2). It supports merging, refinement and

reclustering of clusters. It is 'drop in' compatible with the StackPack

clustering package. wcd supports parallelization under both shared memory and

cluster architectures. It is distributed with an EMBOSS wrapper allowing wcd to

be installed as part of an EMBOSS installation (and so provided by a web

server).AVAILABILITY: wcd is distributed under a GPL licence and is available

from http://code.google.com/p/wcdest.

SUPPLEMENTARY INFORMATION: Additional experimental results. The wcd manual, a

companion paper describing underlying algorithms, and all datasets used for

experimentation can also be found at www.bioinf.wits.ac.za/~scott/wcdsupp.html.

DOI: 10.1093/bioinformatics/btn203

PMCID: PMC2718666

PMID: 18480101 [Indexed for MEDLINE]

2401. Bioinformatics. 2008 Jul 1;24(13):1534-5. doi: 10.1093/bioinformatics/btn233.

Epub 2008 May 14.

ProBias: a web-server for the identification of user-specified types of

compositionally biased segments in protein sequences.

Kuznetsov IB(1).

Author information:

(1)Gen\*NY\*sis Center for Excellence in Cancer Genomics, Department of

Epidemiology and Biostatistics, One Discovery Drive, University at Albany,

Rensselaer, NY 12144, USA. IKuznetsov@albany.edu

Most proteins contain compositionally biased segments (CBS) in which one or more

amino acid types are significantly overrepresented. CBS that contain amino acids

with similar chemical properties can have functional and structural importance.

This article describes ProBias, a web-server that searches a protein sequence for

CBS composed of user-specified amino acid types. ProBias utilizes the discrete

scan statistics to estimate statistical significance of CBS and is able to detect

even subtle local deviations from the random independence model. The web-server

also analyzes the global compositional bias of the input sequence. In the case of

novel proteins that lack functional annotation, statistically significant CBS

reported by ProBias can be used to guide the search for potential functionally

important sites or domains.AVAILABILITY: Freely available at

http://lcg.rit.albany.edu/ProBias.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btn233

PMID: 18480099 [Indexed for MEDLINE]

2402. Comput Biol Med. 2008 Jul;38(7):785-91. doi: 10.1016/j.compbiomed.2008.04.005.

Epub 2008 Jun 9.

A web-based tool for the assessment of discrimination and calibration properties

of prognostic models.

Taktak AF(1), Eleuteri A, Lake SP, Fisher AC.

Author information:

(1)Department of Clinical Engineering, Royal Liverpool University Hospital, 1st

Floor, Duncan Building, Daulby Street, Liverpool L7 8XP, UK. afgt@liv.ac.uk

Prognostic models are developed to assist clinicians in making decisions

regarding treatment and follow-up management. The accuracy of these models is

often assessed either in terms of their discrimination performance or calibration

but rarely both. In this paper, we describe the development of an online tool for

discrimination using Harrell C index and calibration using a Hosmer-Lemeshow type

analysis (http://clinengnhs.liv.ac.uk/AADP/AADP\_Welcome.htm). We show examples of

using the tool on real data. We highlight situations where the model performed

well in terms of either discrimination or calibration but not both depending on

the sample size of the test set. We conclude that prognostic models should be

assessed both in terms of discrimination and calibration and that calibration

analysis should be carried out numerically and graphically.

DOI: 10.1016/j.compbiomed.2008.04.005

PMID: 18539267 [Indexed for MEDLINE]

2403. J Chem Inf Model. 2008 Jul;48(7):1524-9. doi: 10.1021/ci8000474. Epub 2008 Jun

21.

Web server to identify similarity of amino acid motifs to compounds (SAAMCO).

Casey FP(1), Davey NE, Baran I, Varekova RS, Shields DC.

Author information:

(1)UCD Conway Institute of Biomolecular and Biomedical Sciences and Complex and

Adaptive Systems Laboratory (CASL), University College Dublin, Dublin, Ireland.

Protein-protein interactions are fundamental in mediating biological processes

including metabolism, cell growth, and signaling. To be able to selectively

inhibit or induce protein activity or complex formation is a key feature in

controlling disease. For those situations in which protein-protein interactions

derive substantial affinity from short linear peptide sequences, or motifs, we

can develop search algorithms for peptidomimetic compounds that resemble the

short peptide's structure but are not compromised by poor pharmacological

properties. SAAMCO is a Web service ( http://bioware.ucd.ie/ approximately

saamco) that facilitates the screening of motifs with known structures against

bioactive compound databases. It is built on an algorithm that defines compound

similarity based on the presence of appropriate amino acid side chain fragments

and a favorable Root Mean Squared Deviation (RMSD) between compound and motif

structure. The methodology is efficient as the available compound databases are

preprocessed and fast regular expression searches filter potential matches before

time-intensive 3D superposition is performed. The required input information is

minimal, and the compound databases have been selected to maximize the

availability of information on biological activity. "Hits" are accompanied with a

visualization window and links to source database entries. Motif matching can be

defined on partial or full similarity which will increase or reduce respectively

the number of potential mimetic compounds. The Web server provides the

functionality for rapid screening of known or putative interaction motifs against

prepared compound libraries using a novel search algorithm. The tabulated results

can be analyzed by linking to appropriate databases and by visualization.

DOI: 10.1021/ci8000474

PMID: 18570372 [Indexed for MEDLINE]

2404. J Environ Monit. 2008 Jul;10(7):812-6. doi: 10.1039/b719165k. Epub 2008 Jun 17.

Large scale remote sensing for environmental monitoring of infrastructure.

Whelan MJ(1), Fuchs MP, Janoyan KD.

Author information:

(1)Clarkson University, Department of Civil and Environmental Engineering,

Potsdam, NY13699, USA. whelanmj@clarkson.edu

Recent developments in wireless sensor technology afford the opportunity to

rapidly and easily deploy large-scale, low-cost, and low-power sensor networks

across relatively sizeable environmental regions. Furthermore, the advancement of

increasingly smaller and less expensive wireless hardware is further complemented

by the rapid development of open-source software components. These software

protocols allow for interfacing with the hardware to program and configure the

onboard processing and communication settings. In general, a wireless sensor

network topology consists of an array of microprocessor boards, referred to as

motes, which can engage in two-way communication among each other as well as with

a base station that relays the mote data to a host computer. The information can

then be either logged and displayed on the local host or directed to an http

server for network monitoring remote from the site. A number of wireless sensor

products are available that offer off-the-shelf network hardware as well as

sensor solutions for environmental monitoring that are compatible with the TinyOS

open-source software platform. This paper presents an introduction to wireless

sensing and to the use of external antennas for increasing the antenna radiation

intensity and shaping signal directivity for monitoring applications requiring

larger mote-to-mote communication distances.

DOI: 10.1039/b719165k

PMID: 18688447 [Indexed for MEDLINE]

2405. J Mol Recognit. 2008 Jul-Aug;21(4):243-55. doi: 10.1002/jmr.893.

Predicting linear B-cell epitopes using string kernels.

El-Manzalawy Y(1), Dobbs D, Honavar V.

Author information:

(1)Artificial Intelligence Laboratory, Iowa State University, Ames, IA 50010,

USA. yasser@iastate.edu

The identification and characterization of B-cell epitopes play an important role

in vaccine design, immunodiagnostic tests, and antibody production. Therefore,

computational tools for reliably predicting linear B-cell epitopes are highly

desirable. We evaluated Support Vector Machine (SVM) classifiers trained

utilizing five different kernel methods using fivefold cross-validation on a

homology-reduced data set of 701 linear B-cell epitopes, extracted from Bcipep

database, and 701 non-epitopes, randomly extracted from SwissProt sequences.

Based on the results of our computational experiments, we propose BCPred, a novel

method for predicting linear B-cell epitopes using the subsequence kernel. We

show that the predictive performance of BCPred (AUC = 0.758) outperforms 11

SVM-based classifiers developed and evaluated in our experiments as well as our

implementation of AAP (AUC = 0.7), a recently proposed method for predicting

linear B-cell epitopes using amino acid pair antigenicity. Furthermore, we

compared BCPred with AAP and ABCPred, a method that uses recurrent neural

networks, using two data sets of unique B-cell epitopes that had been previously

used to evaluate ABCPred. Analysis of the data sets used and the results of this

comparison show that conclusions about the relative performance of different

B-cell epitope prediction methods drawn on the basis of experiments using data

sets of unique B-cell epitopes are likely to yield overly optimistic estimates of

performance of evaluated methods. This argues for the use of carefully

homology-reduced data sets in comparing B-cell epitope prediction methods to

avoid misleading conclusions about how different methods compare to each other.

Our homology-reduced data set and implementations of BCPred as well as the APP

method are publicly available through our web-based server, BCPREDS, at:

http://ailab.cs.iastate.edu/bcpreds/.

John Wiley & Sons, Ltd

DOI: 10.1002/jmr.893

PMCID: PMC2683948

PMID: 18496882 [Indexed for MEDLINE]

2406. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W2-4. doi: 10.1093/nar/gkn399.

Keeping pace with the data: 2008 update on the Bioinformatics Links Directory.

Brazas MD(1), Fox JA, Brown T, McMillan S, Ouellette BF.

Author information:

(1)Ontario Institute for Cancer Research, 101 College St, Suite 800, Toronto,

Ontario, Canada.

The Bioinformatics Links Directory, http://bioinformatics.ca/links\_directory/, is

an online resource for public access to all of the life science research web

servers published in this and previous issues of Nucleic Acids Research, together

with other useful tools, databases and resources for bioinformatics and molecular

biology research. Dependent on community input and development, the

Bioinformatics Links Directory exemplifies an open access research tool and

resource. The 2008 update includes the 94 web servers featured in the July 2008

Web Server issue of Nucleic Acids Research, bringing the total number of servers

listed in the Bioinformatics Links Directory to over 1200 links. A complete list

of all links listed in this Nucleic Acids Research 2008 Web Server issue can be

accessed online at http://bioinfomatics.ca/links\_directory/narweb2008/. The 2008

update of the Bioinformatics Links Directory, which includes the Web Server list

and summaries, is also available online at the Nucleic Acids Research website,

http://nar.oxfordjournals.org/.

DOI: 10.1093/nar/gkn399

PMCID: PMC2447757

PMID: 18586831 [Indexed for MEDLINE]

2407. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W216-22. doi:

10.1093/nar/gkn367. Epub 2008 Jun 13.

LocalMove: computing on-lattice fits for biopolymers.

Ponty Y(1), Istrate R, Porcelli E, Clote P.

Author information:

(1)Department of Biology, Boston College, Chestnut Hill, MA 02467, USA.

Given an input Protein Data Bank file (PDB) for a protein or RNA molecule,

LocalMove is a web server that determines an on-lattice representation for the

input biomolecule. The web server implements a Markov Chain Monte-Carlo algorithm

with simulated annealing to compute an approximate fit for either the

coarse-grain model or backbone model on either the cubic or face-centered cubic

lattice. LocalMove returns a PDB file as output, as well as dynamic movie of 3D

images of intermediate conformations during the computation. The LocalMove server

is publicly available at http://bioinformatics.bc.edu/clotelab/localmove/.

DOI: 10.1093/nar/gkn367

PMCID: PMC2447748

PMID: 18556754 [Indexed for MEDLINE]

2408. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W352-7. doi:

10.1093/nar/gkn323. Epub 2008 Jun 10.

Onto-CC: a web server for identifying Gene Ontology conceptual clusters.

Romero-Zaliz R(1), Del Val C, Cobb JP, Zwir I.

Author information:

(1)Departamento de Ciencias de la Computación e Inteligencia Artificial, Escuela

Técnica Superior de Ingenierías Informática y de Telecomunicación, c/. Daniel

Saucedo Aranda, s/n 18071 Granada, Spain.

The Gene Ontology (GO) vocabulary has been extensively explored to analyze the

functions of coexpressed genes. However, despite its extended use in Biology and

Medical Sciences, there are still high levels of uncertainty about which ontology

(i.e. Molecular Process, Cellular Component or Molecular Function) should be

used, and at which level of specificity. Moreover, the GO database can contain

incomplete information resulting from human annotations, or highly influenced by

the available knowledge about a specific branch in an ontology. In spite of these

drawbacks, there is a trend to ignore these problems and even use GO terms to

conduct searches of gene expression profiles (i.e. expression + GO) instead of

more cautious approaches that just consider them as an independent source of

validation (i.e. expression versus GO). Consequently, propagating the uncertainty

and producing biased analysis of the required gene grouping hypotheses. We

proposed a web tool, Onto-CC, as an automatic method specially suited for

independent explanation/validation of gene grouping hypotheses (e.g. coexpressed

genes) based on GO clusters (i.e. expression versus GO). Onto-CC approach reduces

the uncertainty of the queries by identifying optimal conceptual clusters that

combine terms from different ontologies simultaneously, as well as terms defined

at different levels of specificity in the GO hierarchy. To do so, we implemented

the EMO-CC methodology to find clusters in structural databases [GO Directed

acyclic Graph (DAG) tree], inspired on Conceptual Clustering algorithms. This

approach allows the management of optimal cluster sets as potential parallel

hypotheses, guided by multiobjective/multimodal optimization techniques.

Therefore, we can generate alternative and, still, optimal explanations of

queries that can provide new insights for a given problem. Onto-CC has been

successfully used to test different medical and biological hypotheses including

the explanation and prediction of gene expression profiles resulting from the

host response to injuries in the inflammatory problem. Onto-CC provides two

versions: Ready2GO, a precalculated EMO-CC for several genomes and an Advanced

Onto-CC for custom annotation files

(http://gps-tools2.wustl.edu/onto-cc/index.html).

DOI: 10.1093/nar/gkn323

PMCID: PMC2447763

PMID: 18544607 [Indexed for MEDLINE]

2409. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W433-7. doi:

10.1093/nar/gkn284. Epub 2008 Jun 6.

Comparative Pathway Analyzer--a web server for comparative analysis, clustering

and visualization of metabolic networks in multiple organisms.

Oehm S(1), Gilbert D, Tauch A, Stoye J, Goesmann A.

Author information:

(1)Center for Biotechnology (CeBiTec), Bielefeld University, D-33594 Bielefeld,

Germany. sebastian.oehm@cebitec.uni-bietefeld.de

In order to understand the phenotype of any living system, it is essential to not

only investigate its genes, but also the specific metabolic pathway variant of

the organism of interest, ideally in comparison with other organisms. The

Comparative Pathway Analyzer, CPA, calculates and displays the differences in

metabolic reaction content between two sets of organisms. Because results are

highly dependent on the distribution of organisms into these two sets and the

appropriate definition of these sets often is not easy, we provide hierarchical

clustering methods for the identification of significant groupings. CPA also

visualizes the reaction content of several organisms simultaneously allowing easy

comparison. Reaction annotation data and maps for visualizing the results are

taken from the KEGG database. Additionally, users can upload their own annotation

data. This website is free and open to all users and there is no login

requirement. It is available at

https://www.cebitec.uni-bielefeld.de/groups/brf/software/cpa/index.html.

DOI: 10.1093/nar/gkn284

PMCID: PMC2447754

PMID: 18539612 [Indexed for MEDLINE]

2410. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W265-9. doi:

10.1093/nar/gkn346. Epub 2008 Jun 6.

KFC Server: interactive forecasting of protein interaction hot spots.

Darnell SJ(1), LeGault L, Mitchell JC.

Author information:

(1)Department of Biochemistry, University of Wisconsin-Madison, Madison, WI

53706, USA.

The KFC Server is a web-based implementation of the KFC (Knowledge-based FADE and

Contacts) model-a machine learning approach for the prediction of binding hot

spots, or the subset of residues that account for most of a protein interface's;

binding free energy. The server facilitates the automated analysis of a user

submitted protein-protein or protein-DNA interface and the visualization of its

hot spot predictions. For each residue in the interface, the KFC Server

characterizes its local structural environment, compares that environment to the

environments of experimentally determined hot spots and predicts if the interface

residue is a hot spot. After the computational analysis, the user can visualize

the results using an interactive job viewer able to quickly highlight predicted

hot spots and surrounding structural features within the protein structure. The

KFC Server is accessible at http://kfc.mitchell-lab.org.

DOI: 10.1093/nar/gkn346

PMCID: PMC2447760

PMID: 18539611 [Indexed for MEDLINE]

2411. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W385-9. doi:

10.1093/nar/gkn317. Epub 2008 May 31.

BioLit: integrating biological literature with databases.

Fink JL(1), Kushch S, Williams PR, Bourne PE.

Author information:

(1)Skaggs School of Pharmacy and Pharmaceutical Sciences, University of

California, San Diego, CA, 92093-0444, USA.

BioLit is a web server which provides metadata describing the semantic content of

all open access, peer-reviewed articles which describe research from the major

life sciences literature archive, PubMed Central. Specifically, these metadata

include database identifiers and ontology terms found within the full text of the

article. BioLit delivers these metadata in the form of XML-based article files

and as a custom web-based article viewer that provides context-specific

functionality to the metadata. This resource aims to integrate the traditional

scientific publication directly into existing biological databases, thus

obviating the need for a user to search in multiple locations for information

relating to a specific item of interest, for example published experimental

results associated with a particular biological database entry. As an example of

a possible use of BioLit, we also present an instance of the Protein Data Bank

fully integrated with BioLit data. We expect that the community of life

scientists in general will be the primary end-users of the web-based viewer,

while biocurators will make use of the metadata-containing XML files and the

BioLit database of article data. BioLit is available at http://biolit.ucsd.edu.

DOI: 10.1093/nar/gkn317

PMCID: PMC2447735

PMID: 18515836 [Indexed for MEDLINE]

2412. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W496-502. doi:

10.1093/nar/gkn305. Epub 2008 May 30.

CS23D: a web server for rapid protein structure generation using NMR chemical

shifts and sequence data.

Wishart DS(1), Arndt D, Berjanskii M, Tang P, Zhou J, Lin G.

Author information:

(1)Department of Computing Science, Department of Biological Sciences, University

of Alberta and National Research Council, National Institute for Nanotechnology

(NINT), Edmonton, AB, Canada T6G 2E8.

CS23D (chemical shift to 3D structure) is a web server for rapidly generating

accurate 3D protein structures using only assigned nuclear magnetic resonance

(NMR) chemical shifts and sequence data as input. Unlike conventional NMR

methods, CS23D requires no NOE and/or J-coupling data to perform its

calculations. CS23D accepts chemical shift files in either SHIFTY or BMRB

formats, and produces a set of PDB coordinates for the protein in about 10-15

min. CS23D uses a pipeline of several preexisting programs or servers to

calculate the actual protein structure. Depending on the sequence similarity (or

lack thereof) CS23D uses either (i) maximal subfragment assembly (a form of

homology modeling), (ii) chemical shift threading or (iii) shift-aided de novo

structure prediction (via Rosetta) followed by chemical shift refinement to

generate and/or refine protein coordinates. Tests conducted on more than 100

proteins from the BioMagResBank indicate that CS23D converges (i.e. finds a

solution) for >95% of protein queries. These chemical shift generated structures

were found to be within 0.2-2.8 A RMSD of the NMR structure generated using

conventional NOE-base NMR methods or conventional X-ray methods. The performance

of CS23D is dependent on the completeness of the chemical shift assignments and

the similarity of the query protein to known 3D folds. CS23D is accessible at

http://www.cs23d.ca.

DOI: 10.1093/nar/gkn305

PMCID: PMC2447725

PMID: 18515350 [Indexed for MEDLINE]

2413. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W303-7. doi:

10.1093/nar/gkn308. Epub 2008 May 30.

3D-Fun: predicting enzyme function from structure.

von Grotthuss M(1), Plewczynski D, Vriend G, Rychlewski L.

Author information:

(1)CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Geert Grooteplein

26-28, 6525 GA Nijmegen, The Netherlands.

The 'omics' revolution is causing a flurry of data that all needs to be annotated

for it to become useful. Sequences of proteins of unknown function can be

annotated with a putative function by comparing them with proteins of known

function. This form of annotation is typically performed with BLAST or similar

software. Structural genomics is nowadays also bringing us three dimensional

structures of proteins with unknown function. We present here software that can

be used when sequence comparisons fail to determine the function of a protein

with known structure but unknown function. The software, called 3D-Fun, is

implemented as a server that runs at several European institutes and is freely

available for everybody at all these sites. The 3D-Fun servers accept protein

coordinates in the standard PDB format and compare them with all known protein

structures by 3D structural superposition using the 3D-Hit software. If

structural hits are found with proteins with known function, these are listed

together with their function and some vital comparison statistics. This is

conceptually very similar in 3D to what BLAST does in 1D. Additionally, the

superposition results are displayed using interactive graphics facilities.

Currently, the 3D-Fun system only predicts enzyme function but an expanded

version with Gene Ontology predictions will be available soon. The server can be

accessed at http://3dfun.bioinfo.pl/ or at http://3dfun.cmbi.ru.nl/.

DOI: 10.1093/nar/gkn308

PMCID: PMC2447717

PMID: 18515349 [Indexed for MEDLINE]

2414. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W523-8. doi:

10.1093/nar/gkn335. Epub 2008 May 30.

bioNMF: a web-based tool for nonnegative matrix factorization in biology.

Mejía-Roa E(1), Carmona-Saez P, Nogales R, Vicente C, Vázquez M, Yang XY, García

C, Tirado F, Pascual-Montano A.

Author information:

(1)Computer Architecture Department, Complutense University, Madrid, Spain.

In the last few years, advances in high-throughput technologies are generating

large amounts of biological data that require analysis and interpretation.

Nonnegative matrix factorization (NMF) has been established as a very effective

method to reveal information about the complex latent relationships in

experimental data sets. Using this method as part of the exploratory data

analysis, workflow would certainly help in the process of interpreting and

understanding the complex biology mechanisms that are underlying experimental

data. We have developed bioNMF, a web-based tool that implements the NMF

methodology in different analysis contexts to support some of the most important

reported applications in biology. This online tool provides a user-friendly

interface, combined with a computational efficient parallel implementation of the

NMF methods to explore the data in different analysis scenarios. In addition to

the online access, bioNMF also provides the same functionality included in the

website as a public web services interface, enabling users with more computer

expertise to launch jobs into bioNMF server from their own scripts and workflows.

bioNMF application is freely available at http://bionmf.dacya.ucm.es.

DOI: 10.1093/nar/gkn335

PMCID: PMC2447803

PMID: 18515346 [Indexed for MEDLINE]

2415. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W176-80. doi:

10.1093/nar/gkn330. Epub 2008 May 29.

GeConT 2: gene context analysis for orthologous proteins, conserved domains and

metabolic pathways.

Martinez-Guerrero CE(1), Ciria R, Abreu-Goodger C, Moreno-Hagelsieb G, Merino E.

Author information:

(1)Departmento de Microbiología Molecular, Instituto de Biotecnología,

Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62210, México.

The Gene Context Tool (GeConT) allows users to visualize the genomic context of a

gene or a group of genes and their orthologous relationships within fully

sequenced bacterial genomes. The new version of the server incorporates

information from the COG, Pfam and KEGG databases, allowing users to have an

integrated graphical representation of the function of genes at multiple levels,

their phylogenetic distribution and their genomic context. The sequence of any of

the genes can be easily retrieved, as well as the 5' or 3' regulatory regions,

greatly facilitating further types of analysis. GeConT 2 is available at:

http://bioinfo.ibt.unam.mx/gecont.

DOI: 10.1093/nar/gkn330

PMCID: PMC2447741

PMID: 18511460 [Indexed for MEDLINE]

2416. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W270-5. doi:

10.1093/nar/gkn314. Epub 2008 May 28.

PBEQ-Solver for online visualization of electrostatic potential of biomolecules.

Jo S(1), Vargyas M, Vasko-Szedlar J, Roux B, Im W.

Author information:

(1)Department of Chemistry, The University of Kansas, 2030 Becker Drive,

Lawrence, KS 66047, USA.

PBEQ-Solver provides a web-based graphical user interface to read biomolecular

structures, solve the Poisson-Boltzmann (PB) equations and interactively

visualize the electrostatic potential. PBEQ-Solver calculates (i) electrostatic

potential and solvation free energy, (ii) protein-protein (DNA or RNA)

electrostatic interaction energy and (iii) pKa of a selected titratable residue.

All the calculations can be performed in both aqueous solvent and membrane

environments (with a cylindrical pore in the case of membrane). PBEQ-Solver uses

the PBEQ module in the biomolecular simulation program CHARMM to solve the

finite-difference PB equation of molecules specified by users. Users can

interactively inspect the calculated electrostatic potential on the

solvent-accessible surface as well as iso-electrostatic potential contours using

a novel online visualization tool based on MarvinSpace molecular visualization

software, a Java applet integrated within CHARMM-GUI (http://www.charmm-gui.org).

To reduce the computational time on the server, and to increase the efficiency in

visualization, all the PB calculations are performed with coarse grid spacing

(1.5 A before and 1 A after focusing). PBEQ-Solver suggests various physical

parameters for PB calculations and users can modify them if necessary.

PBEQ-Solver is available at http://www.charmm-gui.org/input/pbeqsolver.

DOI: 10.1093/nar/gkn314

PMCID: PMC2447802

PMID: 18508808 [Indexed for MEDLINE]

2417. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W25-9. doi:

10.1093/nar/gkn320. Epub 2008 May 24.

SCANPS: a web server for iterative protein sequence database searching by dynamic

programing, with display in a hierarchical SCOP browser.

Walsh TP(1), Webber C, Searle S, Sturrock SS, Barton GJ.

Author information:

(1)College of Life Sciences, University of Dundee, Dundee, UK.

SCANPS performs iterative profile searching similar to PSI-BLAST but with full

dynamic programing on each cycle and on-the-fly estimation of significance. This

combination gives good sensitivity and selectivity that outperforms PSI-BLAST in

domain-searching benchmarks. Although computationally expensive, SCANPS exploits

onchip parallelism (MMX and SSE2 instructions on Intel chips) as well as MPI

parallelism to give acceptable turnround times even for large databases. A web

server developed to run SCANPS searches is now available at

http://www.compbio.dundee.ac.uk/www-scanps. The server interface allows a range

of different protein sequence databases to be searched including the SCOP

database of protein domains. The server provides the user with regularly updated

versions of the main protein sequence databases and is backed up by significant

computing resources which ensure that searches are performed rapidly. For SCOP

searches, the results may be viewed in a new tree-based representation that

reflects the structure of the SCOP hierarchy; this aids the user in placing each

hit in the context of its SCOP classification and understanding its relationship

to other domains in SCOP.

DOI: 10.1093/nar/gkn320

PMCID: PMC2447745

PMID: 18503088 [Indexed for MEDLINE]

2418. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W30-4. doi:

10.1093/nar/gkn322. Epub 2008 May 24.

PROMALS3D web server for accurate multiple protein sequence and structure

alignments.

Pei J(1), Tang M, Grishin NV.

Author information:

(1)Howard Hughes Medical Institute and Department of Biochemistry, University of

Texas Southwestern Medical Center, 6001 Forest Park Road, Dallas, TX 75390-9050,

USA.

Multiple sequence alignments are essential in computational sequence and

structural analysis, with applications in homology detection, structure modeling,

function prediction and phylogenetic analysis. We report PROMALS3D web server for

constructing alignments for multiple protein sequences and/or structures using

information from available 3D structures, database homologs and predicted

secondary structures. PROMALS3D shows higher alignment accuracy than a number of

other advanced methods. Input of PROMALS3D web server can be FASTA format protein

sequences, PDB format protein structures and/or user-defined alignment

constraints. The output page provides alignments with several formats, including

a colored alignment augmented with useful information about sequence grouping,

predicted secondary structures and consensus sequences. Intermediate results of

sequence and structural database searches are also available. The PROMALS3D web

server is available at: http://prodata.swmed.edu/promals3d/.

DOI: 10.1093/nar/gkn322

PMCID: PMC2447800

PMID: 18503087 [Indexed for MEDLINE]

2419. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W239-45. doi:

10.1093/nar/gkn326. Epub 2008 May 23.

Domain Hierarchy and closed Loops (DHcL): a server for exploring hierarchy of

protein domain structure.

Koczyk G(1), Berezovsky IN.

Author information:

(1)Computational Biology Unit, Bergen Center for Computational Science,

University of Bergen, Bergen 5008, Norway.

Domain hierarchy and closed loops (DHcL) (http://sitron.bccs.uib.no/dhcl/) is a

web server that delineates energy hierarchy of protein domain structure and

detects domains at different levels of this hierarchy. The server also identifies

closed loops and van der Waals locks, which constitute a structural basis for the

protein domain hierarchy. The DHcL can be a useful tool for an express analysis

of protein structures and their alternative domain decompositions. The user

submits a PDB identifier(s) or uploads a 3D protein structure in a PDB format.

The results of the analysis are the location of domains at different levels of

hierarchy, closed loops, van der Waals locks and their interactive visualization.

The server maintains a regularly updated database of domains, closed loop and van

der Waals locks for all X-ray structures in PDB. DHcL server is available at:

http://sitron.bccs.uib.no/dhcl.

DOI: 10.1093/nar/gkn326

PMCID: PMC2447749

PMID: 18502776 [Indexed for MEDLINE]

2420. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W55-9. doi:

10.1093/nar/gkn307. Epub 2008 May 22.

SuperPred: drug classification and target prediction.

Dunkel M(1), Günther S, Ahmed J, Wittig B, Preissner R.

Author information:

(1)Institute of Molecular Biology and Bioinformatics, Charité - University

Medicine Berlin, Arnimallee 22, 14195 Berlin, Germany.

The drug classification scheme of the World Health Organization (WHO) [Anatomical

Therapeutic Chemical (ATC)-code] connects chemical classification and therapeutic

approach. It is generally accepted that compounds with similar physicochemical

properties exhibit similar biological activity. If this hypothesis holds true for

drugs, then the ATC-code, the putative medical indication area and potentially

the medical target should be predictable on the basis of structural similarity.

We have validated that the prediction of the drug class is reliable for

WHO-classified drugs. The reliability of the predicted medical effects of the

compounds increases with a rising number of (physico-) chemical properties

similar to a drug with known function. The web-server translates a user-defined

molecule into a structural fingerprint that is compared to about 6300 drugs,

which are enriched by 7300 links to molecular targets of the drugs, derived

through text mining followed by manual curation. Links to the affected pathways

are provided. The similarity to the medical compounds is expressed by the

Tanimoto coefficient that gives the structural similarity of two compounds. A

similarity score higher than 0.85 results in correct ATC prediction for 81% of

all cases. As the biological effect is well predictable, if the structural

similarity is sufficient, the web-server allows prognoses about the medical

indication area of novel compounds and to find new leads for known

targets.AVAILABILITY: the system is freely accessible at

http://bioinformatics.charite.de/superpred. SuperPred can be obtained via a

Creative Commons Attribution Noncommercial-Share Alike 3.0 License.

DOI: 10.1093/nar/gkn307

PMCID: PMC2447784

PMID: 18499712 [Indexed for MEDLINE]

2421. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W119-27. doi:

10.1093/nar/gkn304. Epub 2008 May 21.

RSAT: regulatory sequence analysis tools.

Thomas-Chollier M(1), Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E,

Brohée S, van Helden J.

Author information:

(1)Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université

Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe, B-1050 Bruxelles,

Belgium.

The regulatory sequence analysis tools (RSAT, http://rsat.ulb.ac.be/rsat/) is a

software suite that integrates a wide collection of modular tools for the

detection of cis-regulatory elements in genome sequences. The suite includes

programs for sequence retrieval, pattern discovery, phylogenetic footprint

detection, pattern matching, genome scanning and feature map drawing. Random

controls can be performed with random gene selections or by generating random

sequences according to a variety of background models (Bernoulli, Markov). Beyond

the original word-based pattern-discovery tools (oligo-analysis and

dyad-analysis), we recently added a battery of tools for matrix-based detection

of cis-acting elements, with some original features (adaptive background models,

Markov-chain estimation of P-values) that do not exist in other matrix-based

scanning tools. The web server offers an intuitive interface, where each program

can be accessed either separately or connected to the other tools. In addition,

the tools are now available as web services, enabling their integration in

programmatic workflows. Genomes are regularly updated from various genome

repositories (NCBI and EnsEMBL) and 682 organisms are currently supported. Since

1998, the tools have been used by several hundreds of researchers from all over

the world. Several predictions made with RSAT were validated experimentally and

published.

DOI: 10.1093/nar/gkn304

PMCID: PMC2447775

PMID: 18495751 [Indexed for MEDLINE]

2422. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W47-54. doi:

10.1093/nar/gkn285. Epub 2008 May 20.

Superimpose: a 3D structural superposition server.

Bauer RA(1), Bourne PE, Formella A, Frömmel C, Gille C, Goede A, Guerler A, Hoppe

A, Knapp EW, Pöschel T, Wittig B, Ziegler V, Preissner R.

Author information:

(1)Charité-Universitätsmedizin Berlin, Structural Bioinformatics Group,

Arnimallee 22, 14195 Berlin, Germany.

The Superimposé webserver performs structural similarity searches with a

preference towards 3D structure-based methods. Similarities can be detected

between small molecules (e.g. drugs), parts of large structures (e.g. binding

sites of proteins) and entire proteins. For this purpose, a number of algorithms

were implemented and various databases are provided. Superimposé assists the user

regarding the selection of a suitable combination of algorithm and database.

After the computation on our server infrastructure, a visual assessment of the

results is provided. The structure-based in silico screening for similar

drug-like compounds enables the detection of scaffold-hoppers with putatively

similar effects. The possibility to find similar binding sites can be of special

interest in the functional analysis of proteins. The search for structurally

similar proteins allows the detection of similar folds with different backbone

topology. The Superimposé server is available at:

http://bioinformatics.charite.de/superimpose.

DOI: 10.1093/nar/gkn285

PMCID: PMC2447795

PMID: 18492720 [Indexed for MEDLINE]

2423. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W104-8. doi:

10.1093/nar/gkn250. Epub 2008 May 19.

OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics.

Lu ZJ(1), Mathews DH.

Author information:

(1)Department of Biochemistry and Biophysics, University of Rochester Medical

Center, 601 Elmwood Avenue, Box 712, Rochester, NY 14642, USA.

Given an mRNA sequence as input, the OligoWalk web server generates a list of

small interfering RNA (siRNA) candidate sequences, ranked by the probability of

being efficient siRNA (silencing efficacy greater than 70%). To accomplish this,

the server predicts the free energy changes of the hybridization of an siRNA to a

target mRNA, considering both siRNA and mRNA self-structure. The free energy

changes of the structures are rigorously calculated using a partition function

calculation. By changing advanced options, the free energy changes can also be

calculated using less rigorous lowest free energy structure or suboptimal

structure prediction methods for the purpose of comparison. Considering the

predicted free energy changes and local siRNA sequence features, the server

selects efficient siRNA with high accuracy using a support vector machine. On

average, the fraction of efficient siRNAs selected by the server that will be

efficient at silencing is 78.6%. The OligoWalk web server is freely accessible

through internet at http://rna.urmc.rochester.edu/servers/oligowalk.

DOI: 10.1093/nar/gkn250

PMCID: PMC2447759

PMID: 18490376 [Indexed for MEDLINE]

2424. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W470-4. doi:

10.1093/nar/gkn277. Epub 2008 May 17.

Signature, a web server for taxonomic characterization of sequence samples using

signature genes.

Dutilh BE(1), He Y, Hekkelman ML, Huynen MA.

Author information:

(1)Center for Molecular and Biomolecular Informatics/Nijmegen Center for

Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Geert

Grooteplein 28, 6525 GA, Nijmegen, The Netherlands. dutilh@cmbi.ru.nl

Signature genes are genes that are unique to a taxonomic clade and are common

within it. They contain a wealth of information about clade-specific processes

and hold a strong evolutionary signal that can be used to phylogenetically

characterize a set of sequences, such as a metagenomics sample. As signature

genes are based on gene content, they provide a means to assess the taxonomic

origin of a sequence sample that is complementary to sequence-based analyses.

Here, we introduce Signature (http://www.cmbi.ru.nl/signature), a web server that

identifies the signature genes in a set of query sequences, and therewith

phylogenetically characterizes it. The server produces a list of taxonomic clades

that share signature genes with the set of query sequences, along with an

insightful image of the tree of life, in which the clades are color coded based

on the number of signature genes present. This allows the user to quickly see

from which part(s) of the taxonomy the query sequences likely originate.

DOI: 10.1093/nar/gkn277

PMCID: PMC2447722

PMID: 18487625 [Indexed for MEDLINE]

2425. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W133-9. doi:

10.1093/nar/gkn300. Epub 2008 May 17.

DiRE: identifying distant regulatory elements of co-expressed genes.

Gotea V(1), Ovcharenko I.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.

Regulation of gene expression in eukaryotic genomes is established through a

complex cooperative activity of proximal promoters and distant regulatory

elements (REs) such as enhancers, repressors and silencers. We have developed a

web server named DiRE, based on the Enhancer Identification (EI) method, for

predicting distant regulatory elements in higher eukaryotic genomes, namely for

determining their chromosomal location and functional characteristics. The server

uses gene co-expression data, comparative genomics and profiles of transcription

factor binding sites (TFBSs) to determine TFBS-association signatures that can be

used for discriminating specific regulatory functions. DiRE's unique feature is

its ability to detect REs outside of proximal promoter regions, as it takes

advantage of the full gene locus to conduct the search. DiRE can predict common

REs for any set of input genes for which the user has prior knowledge of

co-expression, co-function or other biologically meaningful grouping. The server

predicts function-specific REs consisting of clusters of specifically-associated

TFBSs and it also scores the association of individual transcription factors

(TFs) with the biological function shared by the group of input genes. Its

integration with the Array2BIO server allows users to start their analysis with

raw microarray expression data. The DiRE web server is freely available at

http://dire.dcode.org.

DOI: 10.1093/nar/gkn300

PMCID: PMC2447744

PMID: 18487623 [Indexed for MEDLINE]

2426. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W170-5. doi:

10.1093/nar/gkn294. Epub 2008 May 16.

QUMA: quantification tool for methylation analysis.

Kumaki Y(1), Oda M, Okano M.

Author information:

(1)Laboratory for Mammalian Epigenetic Studies, Center for Developmental Biology,

RIKEN, 2-2-3 Minatojima-Minamimachi, Chuo-ku, Kobe 650-0047, Japan.

Bisulfite sequencing, a standard method for DNA methylation profile analysis, is

widely used in basic and clinical studies. This method is limited, however, by

the time-consuming data analysis processes required to obtain accurate DNA

methylation profiles from the raw sequence output of the DNA sequencer, and by

the fact that quality checking of the results can be influenced by a researcher's

bias. We have developed an interactive and easy-to-use web-based tool, QUMA

(quantification tool for methylation analysis), for the bisulfite sequencing

analysis of CpG methylation. QUMA includes most of the data-processing functions

necessary for the analysis of bisulfite sequences. It also provides a platform

for consistent quality control of the analysis. The QUMA web server is available

at http://quma.cdb.riken.jp/.

DOI: 10.1093/nar/gkn294

PMCID: PMC2447804

PMID: 18487274 [Indexed for MEDLINE]

2427. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W399-405. doi:

10.1093/nar/gkn296. Epub 2008 May 16.

PolySearch: a web-based text mining system for extracting relationships between

human diseases, genes, mutations, drugs and metabolites.

Cheng D(1), Knox C, Young N, Stothard P, Damaraju S, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Canada.

A particular challenge in biomedical text mining is to find ways of handling

'comprehensive' or 'associative' queries such as 'Find all genes associated with

breast cancer'. Given that many queries in genomics, proteomics or metabolomics

involve these kind of comprehensive searches we believe that a web-based tool

that could support these searches would be quite useful. In response to this

need, we have developed the PolySearch web server. PolySearch supports >50

different classes of queries against nearly a dozen different types of text,

scientific abstract or bioinformatic databases. The typical query supported by

PolySearch is 'Given X, find all Y's' where X or Y can be diseases, tissues, cell

compartments, gene/protein names, SNPs, mutations, drugs and metabolites.

PolySearch also exploits a variety of techniques in text mining and information

retrieval to identify, highlight and rank informative abstracts, paragraphs or

sentences. PolySearch's performance has been assessed in tasks such as gene

synonym identification, protein-protein interaction identification and disease

gene identification using a variety of manually assembled 'gold standard' text

corpuses. Its f-measure on these tasks is 88, 81 and 79%, respectively. These

values are between 5 and 50% better than other published tools. The server is

freely available at http://wishart.biology.ualberta.ca/polysearch.

DOI: 10.1093/nar/gkn296

PMCID: PMC2447794

PMID: 18487273 [Indexed for MEDLINE]

2428. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W202-9. doi:

10.1093/nar/gkn255. Epub 2008 May 15.

PROTEUS2: a web server for comprehensive protein structure prediction and

structure-based annotation.

Montgomerie S(1), Cruz JA, Shrivastava S, Arndt D, Berjanskii M, Wishart DS.

Author information:

(1)Department of Computing Science and Department of Biological Sciences,

University of Alberta and National Research Council, National Institute for

Nanotechnology (NINT), Edmonton, AB, Canada T6G 2E8.

PROTEUS2 is a web server designed to support comprehensive protein structure

prediction and structure-based annotation. PROTEUS2 accepts either single

sequences (for directed studies) or multiple sequences (for whole proteome

annotation) and predicts the secondary and, if possible, tertiary structure of

the query protein(s). Unlike most other tools or servers, PROTEUS2 bundles signal

peptide identification, transmembrane helix prediction, transmembrane beta-strand

prediction, secondary structure prediction (for soluble proteins) and homology

modeling (i.e. 3D structure generation) into a single prediction pipeline. Using

a combination of progressive multi-sequence alignment, structure-based mapping,

hidden Markov models, multi-component neural nets and up-to-date databases of

known secondary structure assignments, PROTEUS is able to achieve among the

highest reported levels of predictive accuracy for signal peptides (Q2 = 94%),

membrane spanning helices (Q2 = 87%) and secondary structure (Q3 score of 81.3%).

PROTEUS2's homology modeling services also provide high quality 3D models that

compare favorably with those generated by SWISS-MODEL and 3D JigSaw (within 0.2 A

RMSD). The average PROTEUS2 prediction takes approximately 3 min per query

sequence. The PROTEUS2 server along with source code for many of its modules is

accessible a http://wishart.biology.ualberta.ca/proteus2.

DOI: 10.1093/nar/gkn255

PMCID: PMC2447806

PMID: 18483082 [Indexed for MEDLINE]

2429. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W10-3. doi:

10.1093/nar/gkn278. Epub 2008 May 15.

R-Coffee: a web server for accurately aligning noncoding RNA sequences.

Moretti S(1), Wilm A, Higgins DG, Xenarios I, Notredame C.

Author information:

(1)Swiss Institute of Bioinformatics (SIB), Quartier Sorge - Genopode, UNIL,

CH-1015 Lausanne, Switzerland.

The R-Coffee web server produces highly accurate multiple alignments of noncoding

RNA (ncRNA) sequences, taking into account predicted secondary structures.

R-Coffee uses a novel algorithm recently incorporated in the T-Coffee package.

R-Coffee works along the same lines as T-Coffee: it uses pairwise or multiple

sequence alignment (MSA) methods to compute a primary library of input

alignments. The program then computes an MSA highly consistent with both the

alignments contained in the library and the secondary structures associated with

the sequences. The secondary structures are predicted using RNAplfold. The server

provides two modes. The slow/accurate mode is restricted to small datasets (less

than 5 sequences less than 150 nucleotides) and combines R-Coffee with Consan, a

very accurate pairwise RNA alignment method. For larger datasets a fast method

can be used (RM-Coffee mode), that uses R-Coffee to combine the output of the

three packages which combines the outputs from programs found to perform best on

RNA (MUSCLE, MAFFT and ProbConsRNA). Our BRAliBase benchmarks indicate that the

R-Coffee/Consan combination is one of the best ncRNA alignment methods for short

sequences, while the RM-Coffee gives comparable results on longer sequences. The

R-Coffee web server is available at http://www.tcoffee.org.

DOI: 10.1093/nar/gkn278

PMCID: PMC2447777

PMID: 18483080 [Indexed for MEDLINE]

2430. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W368-71. doi:

10.1093/nar/gkn256. Epub 2008 May 14.

SerbGO: searching for the best GO tool.

Mosquera JL(1), Sánchez-Pla A.

Author information:

(1)Statistics and Bioinformatics Research Group, Departament d'Estadística,

Universitat de Barcelona. Av. Diagonal 645, 08028 Barcelona, Spain.

jlmosquera@ir.vhebron.net

In recent years, the scientific community has provided many tools to assist with

pathway analysis. Some of these programs can be used to manage functional

annotation of gene products, others are oriented to exploring and analyzing data

sets and many allow both possibilities. Potential users of these tools are faced

with the necessity to decide which of the existing programs are the most

appropriate for their needs. SerbGO is a user-friendly web tool created to

facilitate this task. It can be used (i) to search for specific functionalities

and determine which applications provide them and (ii) to compare several

applications on the basis of different types of functionalities. Iterating and

combining both functionalities can easily lead to selecting an appropriate tool.

Data required by SerbGO is either the desired capabilities within a defined

Standard Functionalities Set or the list of the tools to be compared. The

analysis performed carries out a cross-classification that produces an easily

readable output with the list of tools that implement the capabilities demanded

or a table with the categorization of the GO tools that one wishes to compare.

SerbGO is freely available and does not require a login. It can be accessed

either directly at our server (http://estbioinfo.stat.ub.es/apli/serbgo) or at

the GO Consortium website

(http://www.geneontology.org/GO.tools.microarray.shtml#serbgo).

DOI: 10.1093/nar/gkn256

PMCID: PMC2447766

PMID: 18480123 [Indexed for MEDLINE]

2431. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W97-103. doi:

10.1093/nar/gkn280. Epub 2008 May 14.

AsiDesigner: exon-based siRNA design server considering alternative splicing.

Park YK(1), Park SM, Choi YC, Lee D, Won M, Kim YJ.

Author information:

(1)Medical Genomics Research Center, KRIBB, Daejeon 305-806, Korea.

RNA interference (RNAi) with small interfering RNA (siRNA) has become a powerful

tool in functional and medical genomic research through directed

post-transcriptional gene silencing. In order to apply RNAi technique for

eukaryotic organisms, where frequent alternative splicing results in

diversification of mRNAs and finally of proteins, we need spliced mRNA isoform

silencing to study the function of individual proteins. AsiDesigner is a

web-based siRNA design software system, which provides siRNA design capability to

account for alternative splicing for mRNA level gene silencing. It provides

numerous novel functions including the designing of common siRNAs for the

silencing of more than two mRNAs simultaneously, a scoring scheme to evaluate the

performance of designed siRNAs by adopting currently known key design factors, a

stepwise off-target searching with BLAST and FASTA algorithms and checking the

folding secondary structure energy of siRNAs. To do this, we developed a novel

algorithm to evaluate the common target region, where siRNAs can be designed to

knockdown a specific mRNA isoform or more than two mRNA isoforms from a target

gene simultaneously. The developed algorithm and the AsiDesigner were tested and

validated as very effective throughout widely performed gene silencing

experiments. It is expected that AsiDesigner will play an important role in

functional genomics, drug discovery and other molecular biological research.

AsiDesigner is freely accessible at http://sysbio.kribb.re.kr/AsiDesigner/.

DOI: 10.1093/nar/gkn280

PMCID: PMC2447810

PMID: 18480122 [Indexed for MEDLINE]

2432. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W332-5. doi:

10.1093/nar/gkn289. Epub 2008 May 14.

MADNet: microarray database network web server.

Segota I(1), Bartonicek N, Vlahovicek K.

Author information:

(1)Bioinformatics Group, Division of Biology, Faculty of Science, Zagreb

University, Horvatovac 102a, 10000 Zagreb, Croatia.

MADNet is a user-friendly data mining and visualization tool for rapid analysis

of diverse high-throughput biological data such as microarray, phage display or

even metagenome experiments. It presents biological information in the context of

metabolic and signalling pathways, transcription factors and drug targets through

minimal user input, consisting only of the file with the experimental data. These

data are integrated with information stored in various biological databases such

as NCBI nucleotide and protein databases, metabolic and signalling pathway

databases (KEGG), transcription regulation (TRANSFAC(c)) and drug target database

(DrugBank). MADNet is freely available for academic use at

http://www.bioinfo.hr/madnet.

DOI: 10.1093/nar/gkn289

PMCID: PMC2447778

PMID: 18480121 [Indexed for MEDLINE]

2433. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W286-90. doi:

10.1093/nar/gkn279. Epub 2008 May 13.

The Predikin webserver: improved prediction of protein kinase peptide specificity

using structural information.

Saunders NF(1), Kobe B.

Author information:

(1)School of Molecular and Microbial Sciences, University of Queensland, Brisbane

4072, Australia. n.saunders@uq.edu.au

The Predikin webserver allows users to predict substrates of protein kinases. The

Predikin system is built from three components: a database of protein kinase

substrates that links phosphorylation sites with specific protein kinase

sequences; a perl module to analyse query protein kinases and a web interface

through which users can submit protein kinases for analysis. The Predikin perl

module provides methods to (i) locate protein kinase catalytic domains in a

sequence, (ii) classify them by type or family, (iii) identify

substrate-determining residues, (iv) generate weighted scoring matrices using

three different methods, (v) extract putative phosphorylation sites in query

substrate sequences and (vi) score phosphorylation sites for a given kinase,

using optional filters. The web interface provides user-friendly access to each

of these functions and allows users to obtain rapidly a set of predictions that

they can export for further analysis. The server is available at

http://predikin.biosci.uq.edu.au.

DOI: 10.1093/nar/gkn279

PMCID: PMC2447752

PMID: 18477637 [Indexed for MEDLINE]

2434. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W246-51. doi:

10.1093/nar/gkn259. Epub 2008 May 13.

pFlexAna: detecting conformational changes in remotely related proteins.

Nigham A(1), Tucker-Kellogg L, Mihalek I, Verma C, Hsu D.

Author information:

(1)Department of Computer Science, National University of Singapore, Singapore

117590, Singapore-MIT Alliance, Singapore.

The pFlexAna (protein flexibility analyzer) web server detects and displays

conformational changes in remotely related proteins, without relying on sequence

homology. To do so, it first applies a reliable statistical test to align core

protein fragments that are structurally similar and then clusters these aligned

fragment pairs into 'super-alignments', according to the similarity of geometric

transformations that align them. The result is that the dominant conformational

changes occur between the clusters, while the smaller conformational changes

occur within a cluster. pFlexAna is available at

http://bigbird.comp.nus.edu.sg/pfa2/.

DOI: 10.1093/nar/gkn259

PMCID: PMC2447781

PMID: 18477634 [Indexed for MEDLINE]

2435. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W315-9. doi:

10.1093/nar/gkn265. Epub 2008 May 13.

AMIC@: All MIcroarray Clusterings @ once.

Geraci F(1), Pellegrini M, Renda ME.

Author information:

(1)Istituto di Informatica e Telematica del C.N.R., Via Moruzzi 1, Pisa, Italy.

The AMIC@ Web Server offers a light-weight multi-method clustering engine for

microarray gene-expression data. AMIC@ is a highly interactive tool that stresses

user-friendliness and robustness by adopting AJAX technology, thus allowing an

effective interleaved execution of different clustering algorithms and inspection

of results. Among the salient features AMIC@ offers, there are: (i) automatic

file format detection, (ii) suggestions on the number of clusters using a variant

of the stability-based method of Tibshirani et al. (iii) intuitive visual

inspection of the data via heatmaps and (iv) measurements of the clustering

quality using cluster homogeneity. Large data sets can be processed efficiently

by selecting algorithms (such as FPF-SB and k-Boost), specifically designed for

this purpose. In case of very large data sets, the user can opt for a batch-mode

use of the system by means of the Clustering wizard that runs all algorithms at

once and delivers the results via email. AMIC@ is freely available and open to

all users with no login requirement at the following URL

http://bioalgo.iit.cnr.it/amica.

DOI: 10.1093/nar/gkn265

PMCID: PMC2447730

PMID: 18477631 [Indexed for MEDLINE]

2436. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W114-8. doi:

10.1093/nar/gkn297. Epub 2008 May 12.

pssRNAMiner: a plant short small RNA regulatory cascade analysis server.

Dai X(1), Zhao PX.

Author information:

(1)Plant Biology Division, The Samuel Robert Noble Foundation, Ardmore OK 73401,

USA.

In plants, short RNAs including approximately 21-nt microRNA (miRNA) and 21-nt

trans-acting siRNA (ta-siRNA) compose a 'miRNA --> ta-siRNA --> target gene'

cascade pathway that regulates gene expression at the posttranscriptional level.

In this cascade, biogenesis of ta-siRNA clusters requires 21-nt intervals (i.e.

phasing) and miRNA (phase-initiator) cleavage sites on its TAS transcript. Here,

we report a novel web server, pssRNAMiner, which is developed to identify both

the clusters of phased small RNAs as well as the potential phase-initiator. To

detect phased small RNA clusters, the pssRNAMiner maps input small RNAs against

user-specified transcript/genomic sequences, and then identifies phased small RNA

clusters by evaluating P-values of hypergeometric distribution. To identify

potential phase-initiators, pssRNAMiner aligns input phase-initiators with

transcripts of TAS candidates using the Smith-Waterman algorithm. Potential

cleavage sites on TAS candidates are further identified from complementary

regions by weighting the alignment expectation and its distance to detected

phased small RNA clusters. The pssRNAMiner web server is freely available at

http://bioinfo3.noble.org/pssRNAMiner/.

DOI: 10.1093/nar/gkn297

PMCID: PMC2447807

PMID: 18474525 [Indexed for MEDLINE]

2437. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W260-4. doi:

10.1093/nar/gkn185. Epub 2008 May 8.

MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding

sites and their interactions.

Shulman-Peleg A(1), Shatsky M, Nussinov R, Wolfson HJ.

Author information:

(1)School of Computer Science, Raymond and Beverly Sackler Faculty of Exact

Sciences, Tel Aviv University, Tel Aviv 69978, Israel. shulmana@post.ac.il

Analysis of protein-ligand complexes and recognition of spatially conserved

physico-chemical properties is important for the prediction of binding and

function. Here, we present two webservers for multiple alignment and recognition

of binding patterns shared by a set of protein structures. The first webserver,

MultiBind (http://bioinfo3d.cs.tau.ac.il/MultiBind), performs multiple alignment

of protein binding sites. It recognizes the common spatial chemical binding

patterns even in the absence of similarity of the sequences or the folds of the

compared proteins. The input to the MultiBind server is a set of protein-binding

sites defined by interactions with small molecules. The output is a detailed list

of the shared physico-chemical binding site properties. The second webserver,

MAPPIS (http://bioinfo3d.cs.tau.ac.il/MAPPIS), aims to analyze protein-protein

interactions. It performs multiple alignment of protein-protein interfaces

(PPIs), which are regions of interaction between two protein molecules. MAPPIS

recognizes the spatially conserved physico-chemical interactions, which often

involve energetically important hot-spot residues that are crucial for

protein-protein associations. The input to the MAPPIS server is a set of

protein-protein complexes. The output is a detailed list of the shared

interaction properties of the interfaces.

DOI: 10.1093/nar/gkn185

PMCID: PMC2447750

PMID: 18467424 [Indexed for MEDLINE]

2438. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W509-12. doi:

10.1093/nar/gkn202. Epub 2008 May 7.

NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC

class I affinities for peptides of length 8-11.

Lundegaard C(1), Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M.

Author information:

(1)CBS, Department of Systems Biology, Technical University of Denmark DTU,

Kemitorvet Build. 208, 2800 Lyngby, Denmark. lunde@cbc.dtu.dk

NetMHC-3.0 is trained on a large number of quantitative peptide data using both

affinity data from the Immune Epitope Database and Analysis Resource (IEDB) and

elution data from SYFPEITHI. The method generates high-accuracy predictions of

major histocompatibility complex (MHC): peptide binding. The predictions are

based on artificial neural networks trained on data from 55 MHC alleles (43 Human

and 12 non-human), and position-specific scoring matrices (PSSMs) for additional

67 HLA alleles. As only the MHC class I prediction server is available,

predictions are possible for peptides of length 8-11 for all 122 alleles.

artificial neural network predictions are given as actual IC(50) values whereas

PSSM predictions are given as a log-odds likelihood scores. The output is

optionally available as download for easy post-processing. The training method

underlying the server is the best available, and has been used to predict

possible MHC-binding peptides in a series of pathogen viral proteomes including

SARS, Influenza and HIV, resulting in an average of 75-80% confirmed MHC binders.

Here, the performance is further validated and benchmarked using a large set of

newly published affinity data, non-redundant to the training set. The server is

free of use and available at: http://www.cbs.dtu.dk/services/NetMHC.

DOI: 10.1093/nar/gkn202

PMCID: PMC2447772

PMID: 18463140 [Indexed for MEDLINE]

2439. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W255-9. doi:

10.1093/nar/gkn237. Epub 2008 May 7.

SEQATOMS: a web tool for identifying missing regions in PDB in sequence context.

Brandt BW(1), Heringa J, Leunissen JA.

Author information:

(1)Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam, De

Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands.

With over 46 000 proteins, the Protein Data Bank (PDB) is the most important

database with structural information of biological macromolecules. PDB files

contain sequence and coordinate information. Residues present in the sequence can

be absent from the coordinate section, which means their position in space is

unknown. Similarity searches are routinely carried out against sequences taken

from PDB SEQRES. However, there no distinction is made between residues that have

a known or unknown position in the 3D protein structure. We present a FASTA

sequence database that is produced by combining the sequence and coordinate

information. All residues absent from the PDB coordinate section are masked with

lower-case letters, thereby providing a view of these residues in the context of

the entire protein sequence, which facilitates inspecting 'missing' regions. We

also provide a masked version of the CATH domain database. A user-friendly BLAST

interface is available for similarity searching. In contrast to standard

(stand-alone) BLAST output, which only contains upper-case letters, our output

retains the lower-case letters of the masked regions. Thus, our server can be

used to perform BLAST searching case-sensitively. Here, we have applied it to the

study of missing regions in their sequence context. SEQATOMS is available at

http://www.bioinformatics.nl/tools/seqatoms/.

DOI: 10.1093/nar/gkn237

PMCID: PMC2447787

PMID: 18463137 [Indexed for MEDLINE]

2440. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W197-201. doi:

10.1093/nar/gkn238. Epub 2008 May 7.

The Jpred 3 secondary structure prediction server.

Cole C(1), Barber JD, Barton GJ.

Author information:

(1)School of Life Sciences Research, University of Dundee, Dow Street, Dundee,

DD1 5EH, UK.

Jpred (http://www.compbio.dundee.ac.uk/jpred) is a secondary structure prediction

server powered by the Jnet algorithm. Jpred performs over 1000 predictions per

week for users in more than 50 countries. The recently updated Jnet algorithm

provides a three-state (alpha-helix, beta-strand and coil) prediction of

secondary structure at an accuracy of 81.5%. Given either a single protein

sequence or a multiple sequence alignment, Jpred derives alignment profiles from

which predictions of secondary structure and solvent accessibility are made. The

predictions are presented as coloured HTML, plain text, PostScript, PDF and via

the Jalview alignment editor to allow flexibility in viewing and applying the

data. The new Jpred 3 server includes significant usability improvements that

include clearer feedback of the progress or failure of submitted requests.

Functional improvements include batch submission of sequences, summary results

via email and updates to the search databases. A new software pipeline will

enable Jnet/Jpred to continue to be updated in sync with major updates to SCOP

and UniProt and so ensures that Jpred 3 will maintain high-accuracy predictions.

DOI: 10.1093/nar/gkn238

PMCID: PMC2447793

PMID: 18463136 [Indexed for MEDLINE]

2441. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W140-4. doi:

10.1093/nar/gkn253. Epub 2008 May 7.

OREST: the online resource for EST analysis.

Waegele B(1), Schmidt T, Mewes HW, Ruepp A.

Author information:

(1)Institute for Bioinformatics and Systems Biology (MIPS), Helmholtz Zentrum

München - German Research Center for Environmental Health (GmbH), Ingolstädter

Landstrasse 1, D-85764 Neuherberg, Germany.

brigitte.waegele@helmholtz-muenchen.de

The generation of expressed sequence tag (EST) libraries offers an affordable

approach to investigate organisms, if no genome sequence is available. OREST

(http://mips.gsf.de/genre/proj/orest/index.html) is a server-based EST analysis

pipeline, which allows the rapid analysis of large amounts of ESTs or cDNAs from

mammalia and fungi. In order to assign the ESTs to genes or proteins OREST maps

DNA sequences to reference datasets of gene products and in a second step to

complete genome sequences. Mapping against genome sequences recovers additional

13% of EST data, which otherwise would escape further analysis. To enable

functional analysis of the datasets, ESTs are functionally annotated using the

hierarchical FunCat annotation scheme as well as GO annotation terms. OREST also

allows to predict the association of gene products and diseases by Morbid Map

(OMIM) classification. A statistical analysis of the results of the dataset is

possible with the included PROMPT software, which provides information about

enrichment and depletion of functional and disease annotation terms. OREST was

successfully applied for the identification and functional characterization of

more than 3000 EST sequences of the common marmoset monkey (Callithrix jacchus)

as part of an international collaboration.

DOI: 10.1093/nar/gkn253

PMCID: PMC2447738

PMID: 18463135 [Indexed for MEDLINE]

2442. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W42-6. doi:

10.1093/nar/gkn197. Epub 2008 May 6.

RAPIDO: a web server for the alignment of protein structures in the presence of

conformational changes.

Mosca R(1), Schneider TR.

Author information:

(1)IFOM, the FIRC Institute for Molecular Oncology Foundation, Via Adamello 16,

20139, European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy.

Rapid alignment of proteins in terms of domains (RAPIDO) is a web server for the

3D alignment of crystal structures of different protein molecules in the presence

of conformational change. The structural alignment algorithm identifies groups of

equivalent atoms whose interatomic distances are constant (within a defined

tolerance) in the two structures being compared and considers these groups of

atoms as rigid bodies. In addition to the functionalities provided by existing

tools, RAPIDO can identify structurally equivalent regions also when these

consist of fragments that are distant in terms of sequence and separated by other

movable domains. Furthermore, RAPIDO takes the variation in the reliability of

atomic coordinates into account in the comparison of distances between equivalent

atoms by employing weighting-functions based on the refined B-values. The regions

identified as equivalent by RAPIDO furnish reliable sets of residues for the

superposition of the two structures for subsequent detailed analysis. The RAPIDO

server, with related documentation, is available at

http://webapps.embl-hamburg.de/rapido.

DOI: 10.1093/nar/gkn197

PMCID: PMC2447786

PMID: 18460546 [Indexed for MEDLINE]

2443. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W452-9. doi:

10.1093/nar/gkn230. Epub 2008 May 6.

GraphWeb: mining heterogeneous biological networks for gene modules with

functional significance.

Reimand J(1), Tooming L, Peterson H, Adler P, Vilo J.

Author information:

(1)University of Tartu, Institute of Computer Science, Liivi 2, Tartu, Estonia.

Deciphering heterogeneous cellular networks with embedded modules is a great

challenge of current systems biology. Experimental and computational studies

construct complex networks of molecules that describe various aspects of the cell

such as transcriptional regulation, protein interactions and metabolism. Groups

of interacting genes and proteins reflect network modules that potentially share

regulatory mechanisms and relate to common function. Here, we present GraphWeb, a

public web server for biological network analysis and module discovery. GraphWeb

provides methods to: (1) integrate heterogeneous and multispecies data for

constructing directed and undirected, weighted and unweighted networks; (ii)

discover network modules using a variety of algorithms and topological filters

and (iii) interpret modules using functional knowledge of the Gene Ontology and

pathways, as well as regulatory features such as binding motifs and microRNA

targets. GraphWeb is designed to analyse individual or multiple merged networks,

search for conserved features across multiple species, mine large biological

networks for smaller modules, discover novel candidates and connections for known

pathways and compare results of high-throughput datasets. The GraphWeb is

available at http://biit.cs.ut.ee/graphweb/.

DOI: 10.1093/nar/gkn230

PMCID: PMC2447774

PMID: 18460544 [Indexed for MEDLINE]

2444. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W128-32. doi:

10.1093/nar/gkn195. Epub 2008 May 3.

ConTra: a promoter alignment analysis tool for identification of transcription

factor binding sites across species.

Hooghe B(1), Hulpiau P, van Roy F, De Bleser P.

Author information:

(1)Bioinformatics Core Facility, VIB, Department of Molecular Biology, Ghent

University and Department for Molecular Biomedical Research, VIB, B-9052 Ghent,

Belgium.

Transcription factors (TFs) are key components in signaling pathways, and the

presence of their binding sites in the promoter regions of DNA is essential for

their regulation of the expression of the corresponding genes. Orthologous

promoter sequences are commonly used to increase the specificity with which

potentially functional transcription factor binding sites (TFBSs) are recognized

and to detect possibly important similarities or differences between the

different species. The ConTra (conserved TFBSs) web server provides the biologist

at the bench with a user-friendly tool to interactively visualize TFBSs predicted

using either TransFac (1) or JASPAR (2) position weight matrix libraries, on a

promoter alignment of choice. The visualization can be preceded by a simple

scoring analysis to explore which TFs are the most likely to bind to the promoter

of interest. The ConTra web server is available at

http://bioit.dmbr.ugent.be/ConTra/index.php.

DOI: 10.1093/nar/gkn195

PMCID: PMC2447729

PMID: 18453628 [Indexed for MEDLINE]

2445. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W85-90. doi:

10.1093/nar/gkn220. Epub 2008 Apr 29.

NOBAI: a web server for character coding of geometrical and statistical features

in RNA structure.

Knudsen V(1), Caetano-Anollés G.

Author information:

(1)Department of Crop Sciences, University of Illinois at Urbana-Champaign, IL

61801, USA.

The Numeration of Objects in Biology: Alignment Inferences (NOBAI) web server

provides a web interface to the applications in the NOBAI software package. This

software codes topological and thermodynamic information related to the secondary

structure of RNA molecules as multi-state phylogenetic characters, builds

character matrices directly in NEXUS format and provides sequence randomization

options. The web server is an effective tool that facilitates the search for

evolutionary history embedded in the structure of functional RNA molecules. The

NOBAI web server is accessible at 'http://www.manet.uiuc.edu/nobai/nobai.php'.

This web site is free and open to all users and there is no login requirement.

DOI: 10.1093/nar/gkn220

PMCID: PMC2447726

PMID: 18448469 [Indexed for MEDLINE]

2446. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W210-5. doi:

10.1093/nar/gkn223. Epub 2008 Apr 29.

MolAxis: a server for identification of channels in macromolecules.

Yaffe E(1), Fishelovitch D, Wolfson HJ, Halperin D, Nussinov R.

Author information:

(1)School of Computer Science, Raymond and Beverly Sackler Faculty of Exact

Sciences, Department of Human Genetics, Sackler Institute of Molecular Medicine,

Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel.

MolAxis is a freely available, easy-to-use web server for identification of

channels that connect buried cavities to the outside of macromolecules and for

transmembrane (TM) channels in proteins. Biological channels are essential for

physiological processes such as electrolyte and metabolite transport across

membranes and enzyme catalysis, and can play a role in substrate specificity.

Motivated by the importance of channel identification in macromolecules, we

developed the MolAxis server. MolAxis implements state-of-the-art, accurate

computational-geometry techniques that reduce the dimensions of the channel

finding problem, rendering the algorithm extremely efficient. Given a protein or

nucleic acid structure in the PDB format, the server outputs all possible

channels that connect buried cavities to the outside of the protein or points to

the main channel in TM proteins. For each channel, the gating residues and the

narrowest radius termed 'bottleneck' are also given along with a full list of the

lining residues and the channel surface in a 3D graphical representation. The

users can manipulate advanced parameters and direct the channel search according

to their needs. MolAxis is available as a web server or as a stand-alone program

at http://bioinfo3d.cs.tau.ac.il/MolAxis.

DOI: 10.1093/nar/gkn223

PMCID: PMC2447770

PMID: 18448468 [Indexed for MEDLINE]

2447. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W35-41. doi:

10.1093/nar/gkn211. Epub 2008 Apr 27.

PVS: a web server for protein sequence variability analysis tuned to facilitate

conserved epitope discovery.

Garcia-Boronat M(1), Diez-Rivero CM, Reinherz EL, Reche PA.

Author information:

(1)Immunomedicine Group, Department of Microbiology I, Division of Immunology,

Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense s/n,

Madrid 28040, Spain.

We have developed PVS (Protein Variability Server), a web-based tool that uses

several variability metrics to compute the absolute site variability in multiple

protein-sequence alignments (MSAs). The variability is then assigned to a

user-selected reference sequence consisting of either the first sequence in the

alignment or a consensus sequence. Subsequently, PVS performs tasks that are

relevant for structure-function studies, such as plotting and visualizing the

variability in a relevant 3D-structure. Neatly, PVS also implements some other

tasks that are thought to facilitate the design of epitope discovery-driven

vaccines against pathogens where sequence variability largely contributes to

immune evasion. Thus, PVS can return the conserved fragments in the MSA-as

defined by a user-provided variability threshold-and locate them in a relevant

3D-structure. Furthermore, PVS can return a variability-masked sequence, which

can be directly submitted to the RANKPEP server for the prediction of conserved

T-cell epitopes. PVS is freely available at: http://imed.med.ucm.es/PVS/.

DOI: 10.1093/nar/gkn211

PMCID: PMC2447719

PMID: 18442995 [Indexed for MEDLINE]

2448. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W481-4. doi:

10.1093/nar/gkn194. Epub 2008 Apr 28.

MassTRIX: mass translator into pathways.

Suhre K(1), Schmitt-Kopplin P.

Author information:

(1)Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München,

German Research Center for Environmental Health (GmbH), Ingolstädter Landstrasse

1, 85764 Neuherberg, Germany. karsten.suhre@helmholtz-muenchen.de

Recent technical advances in mass spectrometry (MS) have brought the field of

metabolomics to a point where large numbers of metabolites from numerous

prokaryotic and eukaryotic organisms can now be easily and precisely detected.

The challenge today lies in the correct annotation of these metabolites on the

basis of their accurate measured masses. Assignment of bulk chemical formula is

generally possible, but without consideration of the biological and genomic

context, concrete metabolite annotations remain difficult and uncertain. MassTRIX

responds to this challenge by providing a hypothesis-driven approach to high

precision MS data annotation. It presents the identified chemical compounds in

their genomic context as differentially colored objects on KEGG pathway maps.

Information on gene transcription or differences in the gene complement (e.g.

samples from different bacterial strains) can be easily added. The user can thus

interpret the metabolic state of the organism in the context of its potential

and, in the case of submitted transcriptomics data, real enzymatic capacities.

The MassTRIX web server is freely accessible at http://masstrix.org.

DOI: 10.1093/nar/gkn194

PMCID: PMC2447776

PMID: 18442993 [Indexed for MEDLINE]

2449. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W233-8. doi:

10.1093/nar/gkn216. Epub 2008 Apr 28.

The RosettaDock server for local protein-protein docking.

Lyskov S(1), Gray JJ.

Author information:

(1)Department of Chemical and Biomolecular Engineering and Program in Molecular

and Computational Biophysics, Johns Hopkins University, 3400 N. Charles Street,

Baltimore, MD 21218, USA.

The RosettaDock server (http://rosettadock.graylab.jhu.edu) identifies low-energy

conformations of a protein-protein interaction near a given starting

configuration by optimizing rigid-body orientation and side-chain conformations.

The server requires two protein structures as inputs and a starting location for

the search. RosettaDock generates 1000 independent structures, and the server

returns pictures, coordinate files and detailed scoring information for the 10

top-scoring models. A plot of the total energy of each of the 1000 models created

shows the presence or absence of an energetic binding funnel. RosettaDock has

been validated on the docking benchmark set and through the Critical Assessment

of PRedicted Interactions blind prediction challenge.

DOI: 10.1093/nar/gkn216

PMCID: PMC2447798

PMID: 18442991 [Indexed for MEDLINE]

2450. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W185-9. doi:

10.1093/nar/gkn218. Epub 2008 Apr 28.

HOMCOS: a server to predict interacting protein pairs and interacting sites by

homology modeling of complex structures.

Fukuhara N(1), Kawabata T.

Author information:

(1)Graduate School of Information Science, Nara Institute of Science and

Technology, 8916-5 Takayama, Ikoma, Nara, Japan.

As protein-protein interactions are crucial in most biological processes, it is

valuable to understand how and where protein pairs interact. We developed a web

server HOMCOS (Homology Modeling of Complex Structure,

http://biunit.naist.jp/homcos) to predict interacting protein pairs and

interacting sites by homology modeling of complex structures. Our server is

capable of three services. The first is modeling heterodimers from two query

amino acid sequences posted by users. The server performs BLAST searches to

identify homologous templates in the latest representative dataset of heterodimer

structures generated from the PQS database. Structure validity is evaluated by

the combination of sequence similarity and knowledge-based contact potential

energy as previously described. The server generates a sequence-replaced model

PDB file and a MODELLER script to build full atomic models of complex structures.

The second service is modeling homodimers from one query sequence. The third

service is identification of potentially interacting proteins for one query

sequence. The server searches the dataset of heterodimer structures for a

homologous template, outputs the candidate interacting sequences in the Uniprot

database homologous for the interacting partner template proteins. These features

are useful for wide range of researchers to predict putative interaction sites

and interacting proteins.

DOI: 10.1093/nar/gkn218

PMCID: PMC2447736

PMID: 18442990 [Indexed for MEDLINE]

2451. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W281-5. doi:

10.1093/nar/gkn226. Epub 2008 Apr 28.

GLUE-IT and PEDEL-AA: new programmes for analyzing protein diversity in

randomized libraries.

Firth AE(1), Patrick WM.

Author information:

(1)BioSciences Institute, University College Cork, Cork, Ireland.

There are many methods for introducing random mutations into nucleic acid

sequences. Previously, we described a suite of programmes for estimating the

completeness and diversity of randomized DNA libraries generated by a number of

these protocols. Our programmes suggested some empirical guidelines for library

design; however, no information was provided regarding library diversity at the

protein (rather than DNA) level. We have now updated our web server, enabling

analysis of translated libraries constructed by site-saturation mutagenesis and

error-prone PCR (epPCR). We introduce GLUE-Including Translation (GLUE-IT), which

finds the expected amino acid completeness of libraries in which up to six codons

have been independently varied (according to any user-specified randomization

scheme). We provide two tools for assisting with experimental design:

CodonCalculator, for assessing amino acids corresponding to given randomized

codons; and AA-Calculator, for finding degenerate codons that encode

user-specified sets of amino acids. We also present PEDEL-AA, which calculates

amino acid statistics for libraries generated by epPCR. Input includes the parent

sequence, overall mutation rate, library size, indel rates and a nucleotide

mutation matrix. Output includes amino acid completeness and diversity

statistics, and the number and length distribution of sequences truncated by

premature termination codons. The web interfaces are available at

http://guinevere.otago.ac.nz/stats.html.

DOI: 10.1093/nar/gkn226

PMCID: PMC2447733

PMID: 18442989 [Indexed for MEDLINE]

2452. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W519-22. doi:

10.1093/nar/gkn229. Epub 2008 Apr 24.

EpiToolKit--a web server for computational immunomics.

Feldhahn M(1), Thiel P, Schuler MM, Hillen N, Stevanovic S, Rammensee HG,

Kohlbacher O.

Author information:

(1)Simulation of Biological Systems, Center for Bioinformatics, Institute for

Cell Biology, Eberhard Karls University Tübingen, Germany.

feldhahn@informatik.uni-tuebingen.de

Predicting the T-cell-mediated immune response is an important task in vaccine

design and thus one of the key problems in computational immunomics. Various

methods have been developed during the last decade and are available online. We

present EpiToolKit, a web server that has been specifically designed to offer a

problem-solving environment for computational immunomics. EpiToolKit offers a

variety of different prediction methods for major histocompatibility complex

class I and II ligands as well as minor histocompatibility antigens. These

predictions are embedded in a user-friendly interface allowing refining, editing

and constraining the searches conveniently. We illustrate the value of the

approach with a set of novel tumor-associated peptides. EpiToolKit is available

online at www.epitoolkit.org.

DOI: 10.1093/nar/gkn229

PMCID: PMC2447732

PMID: 18440979 [Indexed for MEDLINE]

2453. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W485-90. doi:

10.1093/nar/gkn196. Epub 2008 Apr 25.

ISPIDER Central: an integrated database web-server for proteomics.

Siepen JA(1), Belhajjame K, Selley JN, Embury SM, Paton NW, Goble CA, Oliver SG,

Stevens R, Zamboulis L, Martin N, Poulovassillis A, Jones P, Côté R, Hermjakob H,

Pentony MM, Jones DT, Orengo CA, Hubbard SJ.

Author information:

(1)Faculty of Life Sciences, University of Manchester, M13 9PT, UK.

Despite the growing volumes of proteomic data, integration of the underlying

results remains problematic owing to differences in formats, data captured,

protein accessions and services available from the individual repositories. To

address this, we present the ISPIDER Central Proteomic Database search

(http://www.ispider.manchester.ac.uk/cgi-bin/ProteomicSearch.pl), an integration

service offering novel search capabilities over leading, mature, proteomic

repositories including PRoteomics IDEntifications database (PRIDE), PepSeeker,

PeptideAtlas and the Global Proteome Machine. It enables users to search for

proteins and peptides that have been characterised in mass spectrometry-based

proteomics experiments from different groups, stored in different databases, and

view the collated results with specialist viewers/clients. In order to overcome

limitations imposed by the great variability in protein accessions used by

individual laboratories, the European Bioinformatics Institute's Protein

Identifier Cross-Reference (PICR) service is used to resolve accessions from

different sequence repositories. Custom-built clients allow users to view

peptide/protein identifications in different contexts from multiple experiments

and repositories, as well as integration with the Dasty2 client supporting any

annotations available from Distributed Annotation System servers. Further

information on the protein hits may also be added via external web services able

to take a protein as input. This web server offers the first truly integrated

access to proteomics repositories and provides a unique service to biologists

interested in mass spectrometry-based proteomics.

DOI: 10.1093/nar/gkn196

PMCID: PMC2447780

PMID: 18440977 [Indexed for MEDLINE]

2454. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W154-6. doi:

10.1093/nar/gkn221. Epub 2008 Apr 25.

OligoHeatMap (OHM): an online tool to estimate and display hybridizations of

oligonucleotides onto DNA sequences.

Croce O(1), Chevenet F, Christen R.

Author information:

(1)University of Nice Sophia-Antipolis and CNRS. UMR 6543, Virtual Biology Lab.

Institute of Developmental Biology and Cancer. Parc Valrose. 06108 Nice, France.

The efficiency of molecular methods involving DNA/DNA hybridizations depends on

the accurate prediction of the melting temperature (T(m)) of the duplex. Many

softwares are available for T(m) calculations, but difficulties arise when one

wishes to check if a given oligomer (PCR primer or probe) hybridizes well or not

on more than a single sequence. Moreover, the presence of mismatches within the

duplex is not sufficient to estimate specificity as it does not always

significantly decrease the T(m). OHM (OligoHeatMap) is an online tool able to

provide estimates of T(m) for a set of oligomers and a set of aligned sequences,

not only as text files of complete results but also in a graphical way: T(m)

values are translated into colors and displayed as a heat map image, either stand

alone or to be used by softwares such as TreeDyn to be included in a phylogenetic

tree. OHM is freely available at http://bioinfo.unice.fr/ohm/, with links to the

full source code and online help.

DOI: 10.1093/nar/gkn221

PMCID: PMC2447727

PMID: 18440971 [Indexed for MEDLINE]

2455. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W75-8. doi:

10.1093/nar/gkn222. Epub 2008 Apr 25.

Software.ncrna.org: web servers for analyses of RNA sequences.

Asai K(1), Kiryu H, Hamada M, Tabei Y, Sato K, Matsui H, Sakakibara Y, Terai G,

Mituyama T.

Author information:

(1)Department of Computational Biology, Graduate School of Frontier Sciences,

University of Tokyo, 5-1-5 Kashiwa-no-ha, Chiba 277-8561, Japan.

We present web servers for analysis of non-coding RNA sequences on the basis of

their secondary structures. Software tools for structural multiple sequence

alignments, structural pairwise sequence alignments and structural motif findings

are available from the integrated web server and the individual stand-alone web

servers. The servers are located at http://software.ncrna.org, along with the

information for the evaluation and downloading. This website is freely available

to all users and there is no login requirement.

DOI: 10.1093/nar/gkn222

PMCID: PMC2447773

PMID: 18440970 [Indexed for MEDLINE]

2456. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W223-8. doi:

10.1093/nar/gkn187. Epub 2008 Apr 19.

PharmaGist: a webserver for ligand-based pharmacophore detection.

Schneidman-Duhovny D(1), Dror O, Inbar Y, Nussinov R, Wolfson HJ.

Author information:

(1)School of Computer Science, Raymond and Beverly Sackler Faculty of Exact

Sciences, Tel Aviv University, Tel Aviv 69978, Israel. duhovka@tau.ac.il

Predicting molecular interactions is a major goal in rational drug design.

Pharmacophore, which is the spatial arrangement of features that is essential for

a molecule to interact with a specific target receptor, is an important model for

achieving this goal. We present a freely available web server, named PharmaGist,

for pharmacophore detection. The employed method is ligand based. Namely, it does

not require the structure of the target receptor. Instead, the input is a set of

structures of drug-like molecules that are known to bind to the receptor. The

output consists of candidate pharmacophores that are computed by multiple

flexible alignment of the input ligands. The method handles the flexibility of

the input ligands explicitly and in deterministic manner within the alignment

process. PharmaGist is also highly efficient, where a typical run with up to 32

drug-like molecules takes seconds to a few minutes on a stardard PC. Another

important characteristic is the capability of detecting pharmacophores shared by

different subsets of input molecules. This capability is a key advantage when the

ligands belong to different binding modes or when the input contains outliers.

The webserver has a user-friendly interface available at

http://bioinfo3d.cs.tau.ac.il/PharmaGist.

DOI: 10.1093/nar/gkn187

PMCID: PMC2447755

PMID: 18424800 [Indexed for MEDLINE]

2457. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W229-32. doi:

10.1093/nar/gkn186. Epub 2008 Apr 19.

FireDock: a web server for fast interaction refinement in molecular docking.

Mashiach E(1), Schneidman-Duhovny D, Andrusier N, Nussinov R, Wolfson HJ.

Author information:

(1)School of Computer Science, Raymond and Beverly Sackler Faculty of Exact

Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

Structural details of protein-protein interactions are invaluable for

understanding and deciphering biological mechanisms. Computational docking

methods aim to predict the structure of a protein-protein complex given the

structures of its single components. Protein flexibility and the absence of

robust scoring functions pose a great challenge in the docking field. Due to

these difficulties most of the docking methods involve a two-tier approach:

coarse global search for feasible orientations that treats proteins as rigid

bodies, followed by an accurate refinement stage that aims to introduce

flexibility into the process. The FireDock web server, presented here, is the

first web server for flexible refinement and scoring of protein-protein docking

solutions. It includes optimization of side-chain conformations and rigid-body

orientation and allows a high-throughput refinement. The server provides a

user-friendly interface and a 3D visualization of the results. A docking protocol

consisting of a global search by PatchDock and a refinement by FireDock was

extensively tested. The protocol was successful in refining and scoring docking

solution candidates for cases taken from docking benchmarks. We provide an option

for using this protocol by automatic redirection of PatchDock candidate solutions

to the FireDock web server for refinement. The FireDock web server is available

at http://bioinfo3d.cs.tau.ac.il/FireDock/.

DOI: 10.1093/nar/gkn186

PMCID: PMC2447790

PMID: 18424796 [Indexed for MEDLINE]

2458. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W70-4. doi:

10.1093/nar/gkn188. Epub 2008 Apr 19.

The Vienna RNA websuite.

Gruber AR(1), Lorenz R, Bernhart SH, Neuböck R, Hofacker IL.

Author information:

(1)Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse

17, 1090 Wien, Austria.

The Vienna RNA Websuite is a comprehensive collection of tools for folding,

design and analysis of RNA sequences. It provides a web interface to the most

commonly used programs of the Vienna RNA package. Among them, we find folding of

single and aligned sequences, prediction of RNA-RNA interactions, and design of

sequences with a given structure. Additionally, we provide analysis of folding

landscapes using the barriers program and structural RNA alignments using

LocARNA. The web server together with software packages for download is freely

accessible at http://rna.tbi.univie.ac.at/.

DOI: 10.1093/nar/gkn188

PMCID: PMC2447809

PMID: 18424795 [Indexed for MEDLINE]

2459. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W276-80. doi:

10.1093/nar/gkn181. Epub 2008 Apr 17.

webPIPSA: a web server for the comparison of protein interaction properties.

Richter S(1), Wenzel A, Stein M, Gabdoulline RR, Wade RC.

Author information:

(1)Molecular and Cellular Modeling Group, EML Research gGmbH,

Schloss-Wolfsbrunnenweg 33, Heidelberg, Germany.

stefan.richter@eml-r.villa-bosch.de

Protein molecular interaction fields are key determinants of protein

functionality. PIPSA (Protein Interaction Property Similarity Analysis) is a

procedure to compare and analyze protein molecular interaction fields, such as

the electrostatic potential. PIPSA may assist in protein functional assignment,

classification of proteins, the comparison of binding properties and the

estimation of enzyme kinetic parameters. webPIPSA is a web server that enables

the use of PIPSA to compare and analyze protein electrostatic potentials. While

PIPSA can be run with downloadable software (see

http://projects.eml.org/mcm/software/pipsa), webPIPSA extends and simplifies a

PIPSA run. This allows non-expert users to perform PIPSA for their protein

datasets. With input protein coordinates, the superposition of protein

structures, as well as the computation and analysis of electrostatic potentials,

is automated. The results are provided as electrostatic similarity matrices from

an all-pairwise comparison of the proteins which can be subjected to clustering

and visualized as epograms (tree-like diagrams showing electrostatic potential

differences) or heat maps. webPIPSA is freely available at: http://pipsa.eml.org.

DOI: 10.1093/nar/gkn181

PMCID: PMC2447742

PMID: 18420653 [Indexed for MEDLINE]

2460. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W60-4. doi:

10.1093/nar/gkn172. Epub 2008 Apr 14.

DAhunter: a web-based server that identifies homologous proteins by comparing

domain architecture.

Lee B(1), Lee D.

Author information:

(1)Korean BioInformation Center, KRIBB, Daejeon 305-806 and Department of Bio and

Brain Engineering, KAIST, Daejeon 305-701, Korea. bulee@kribb.re.kr

We present DAhunter, a web-based server that identifies homologous proteins by

comparing domain architectures, the organization of protein domains. A major

obstacle in comparison of domain architecture is the existence of 'promiscuous'

domains, which carry out auxiliary functions and appear in many unrelated

proteins. To distinguish these promiscuous domains from protein domains, we

assigned a weight score to each domain extracted from RefSeq proteins, based on

its abundance and versatility. A domain's score represents its importance in the

'protein world' and is used in the comparison of domain architectures. In scoring

domains, DAhunter also considers domain combinations as well as single domains.

To measure the similarity of two domain architectures, we developed several

methods that are based on algorithms used in information retrieval (the cosine

similarity, the Goodman-Kruskal gamma function, and domain duplication index) and

then combined these into a similarity score. Compared with other domain

architecture algorithms, DAhunter is better at identifying homology. The server

is available at http://www.dahunter.kr and

http://localodom.kobic.re.kr/dahunter/index.htm.

DOI: 10.1093/nar/gkn172

PMCID: PMC2447808

PMID: 18411203 [Indexed for MEDLINE]

2461. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W181-4. doi:

10.1093/nar/gkn179. Epub 2008 Apr 14.

The CGView Server: a comparative genomics tool for circular genomes.

Grant JR(1), Stothard P.

Author information:

(1)Department of Agricultural, Food and Nutritional Science, University of

Alberta, Canada.

The CGView Server generates graphical maps of circular genomes that show sequence

features, base composition plots, analysis results and sequence similarity plots.

Sequences can be supplied in raw, FASTA, GenBank or EMBL format. Additional

feature or analysis information can be submitted in the form of GFF (General

Feature Format) files. The server uses BLAST to compare the primary sequence to

up to three comparison genomes or sequence sets. The BLAST results and feature

information are converted to a graphical map showing the entire sequence, or an

expanded and more detailed view of a region of interest. Several options are

included to control which types of features are displayed and how the features

are drawn. The CGView Server can be used to visualize features associated with

any bacterial, plasmid, chloroplast or mitochondrial genome, and can aid in the

identification of conserved genome segments, instances of horizontal gene

transfer, and differences in gene copy number. Because a collection of sequences

can be used in place of a comparison genome, maps can also be used to visualize

regions of a known genome covered by newly obtained sequence reads. The CGView

Server can be accessed at http://stothard.afns.ualberta.ca/cgview\_server/

DOI: 10.1093/nar/gkn179

PMCID: PMC2447734

PMID: 18411202 [Indexed for MEDLINE]

2462. Proteins. 2008 Jul;72(1):427-33. doi: 10.1002/prot.21940.

Real-value prediction of backbone torsion angles.

Xue B(1), Dor O, Faraggi E, Zhou Y.

Author information:

(1)Indiana University School of Informatics, Indiana University-Purdue

University, Indianapolis, Indiana 46202, USA.

The backbone structure of a protein is largely determined by the phi and psi

torsion angles. Thus, knowing these angles, even if approximately, will be very

useful for protein-structure prediction. However, in a previous work, a

sequence-based, real-value prediction of psi angle could only achieve a mean

absolute error of 54 degrees (83 degrees, 35 degrees, 33 degrees for coil,

strand, and helix residues, respectively) between predicted and actual angles.

Moreover, a real-value prediction of phi angle is not yet available. This article

employs a neural-network based approach to improve psi prediction by taking

advantage of angle periodicity and apply the new method to the prediction to phi

angles. The 10-fold-cross-validated mean absolute error for the new method is 38

degrees (58 degrees, 33 degrees, 22 degrees for coil, strand, and helix,

respectively) for psi and 25 degrees (35 degrees, 22 degrees, 16 degrees for

coil, strand, and helix, respectively) for phi. The accuracy of real-value

prediction is comparable to or more accurate than the predictions based on

multistate classification of the phi-psi map. More accurate prediction of

real-value angles will likely be useful for improving the accuracy of fold

recognition and ab initio protein-structure prediction. The Real-SPINE 2.0 server

is available on the website http://sparks.informatics.iupui.edu.

2008 Wiley-Liss, Inc.

DOI: 10.1002/prot.21940

PMID: 18214956 [Indexed for MEDLINE]

2463. Zebrafish. 2008 Summer;5(2):125-30. doi: 10.1089/zeb.2008.0531.

FishMap: a community resource for zebrafish genomics.

Meli R(1), Prasad A, Patowary A, Lalwani MK, Maini J, Sharma M, Singh AR, Kumar

G, Jadhav V, Scaria V, Sivasubbu S.

Author information:

(1)Institute of Genomics and Integrative Biology, Council of Scientific and

Industrial Research, Delhi, India.

An enormous amount of information on a genomics scale is available for zebrafish

(Danio rerio), which is a well-studied model organism for human diseases.

However, a majority of this annotation is scattered in obscure data sources.

There have been limited efforts to present it on a unified and integrated

platform, which would help to understand the biological processes in this

organism better. FishMap is a unified and centralized resource for storage,

retrieval, and display of genomic information of zebrafish. The datasets have

been methodically collected from various resources and supplementary information

of publications and mapped to the zebrafish genome. The data are organized into

nine major sections, which include comparative genomics, mapping and sequencing,

gene and gene predictions, expression and regulation, and variation and repeats.

A number of unique sections have been incorporated, which include tracks on

noncoding gene annotation, location of retrovirus/transposon integrations in the

genome, and their flanking genomic sequences and novel transcripts. The datasets

are linked to related data sources. FishMap is built on the Gbrowse, which is a

part of the Generic Model Organism Database Consortium Project. The resource also

features a Web-based BLAST server for sequence homology search and a gene ID

converter that would enable users to sift through different interchangeable gene

annotation identifier systems. The database is amenable to programmatic access

through the Distributed Annotation System as well as BioMoby protocols, thus

making it a central community resource that can be integrated with existing data

mining and analysis workflows. We hope that FishMap would be an integral resource

for community participation in zebrafish genomics. The resource is freely

available at http://miracle.igib.res.in/fishmap, or at

http://fishmap.igib.res.in.

DOI: 10.1089/zeb.2008.0531

PMID: 18554176 [Indexed for MEDLINE]

2464. Bioinformatics. 2008 Jun 15;24(12):1469-70. doi: 10.1093/bioinformatics/btn202.

Epub 2008 Apr 23.

TOPDOM: database of domains and motifs with conservative location in

transmembrane proteins.

Tusnády GE(1), Kalmár L, Hegyi H, Tompa P, Simon I.

Author information:

(1)Institue of Enzymology, BRC, Hungarian Academy of Sciences, H-1113 Karolina út

29, Budapest, Hungary. tusi@enzim.hu

The TOPDOM database is a collection of domains and sequence motifs located

consistently on the same side of the membrane in alpha-helical transmembrane

proteins. The database was created by scanning well-annotated transmembrane

protein sequences in the UniProt database by specific domain or motif detecting

algorithms. The identified domains or motifs were added to the database if they

were uniformly annotated on the same side of the membrane of the various proteins

in the UniProt database. The information about the location of the collected

domains and motifs can be incorporated into constrained topology prediction

algorithms, like HMMTOP, increasing the prediction accuracy.AVAILABILITY: The

TOPDOM database and the constrained HMMTOP prediction server are available on the

page http://topdom.enzim.hu

CONTACT: tusi@enzim.hu; lkalmar@enzim.hu.

DOI: 10.1093/bioinformatics/btn202

PMCID: PMC2427164

PMID: 18434342 [Indexed for MEDLINE]

2465. PLoS One. 2008 Jun 4;3(6):e2325. doi: 10.1371/journal.pone.0002325.

SP5: improving protein fold recognition by using torsion angle profiles and

profile-based gap penalty model.

Zhang W(1), Liu S, Zhou Y.

Author information:

(1)Indiana University School of Informatics and Center for Computational Biology

and Bioinformatics, Indiana University School of Medicine, Indiana

University-Purdue University Indianapolis, Indianapolis, Indiana, United States

of America.

How to recognize the structural fold of a protein is one of the challenges in

protein structure prediction. We have developed a series of single

(non-consensus) methods (SPARKS, SP(2), SP(3), SP(4)) that are based on weighted

matching of two to four sequence and structure-based profiles. There is a robust

improvement of the accuracy and sensitivity of fold recognition as the number of

matching profiles increases. Here, we introduce a new profile-profile comparison

term based on real-value dihedral torsion angles. Together with updated

real-value solvent accessibility profile and a new variable gap-penalty model

based on fractional power of insertion/deletion profiles, the new method (SP(5))

leads to a robust improvement over previous SP method. There is a 2% absolute

increase (5% relative improvement) in alignment accuracy over SP(4) based on two

independent benchmarks. Moreover, SP(5) makes 7% absolute increase (22% relative

improvement) in success rate of recognizing correct structural folds, and 32%

relative improvement in model accuracy of models within the same fold in Lindahl

benchmark. In addition, modeling accuracy of top-1 ranked models is improved by

12% over SP(4) for the difficult targets in CASP 7 test set. These results

highlight the importance of harnessing predicted structural properties in

challenging remote-homolog recognition. The SP(5) server is available at

http://sparks.informatics.iupui.edu.

DOI: 10.1371/journal.pone.0002325

PMCID: PMC2391293

PMID: 18523556 [Indexed for MEDLINE]

2466. BMC Bioinformatics. 2008 Jun 2;9:257. doi: 10.1186/1471-2105-9-257.

SODa: an Mn/Fe superoxide dismutase prediction and design server.

Kwasigroch JM(1), Wintjens R, Gilis D, Rooman M.

Author information:

(1)Unité de Bioinformatique génomique et structurale, Université Libre de

Bruxelles, CP 165/61, avenue Roosevelt 50, B-1050 Brussels, Belgium.

jkwasigr@ulb.ac.be

BACKGROUND: Superoxide dismutases (SODs) are ubiquitous metalloenzymes that play

an important role in the defense of aerobic organisms against oxidative stress,

by converting reactive oxygen species into nontoxic molecules. We focus here on

the SOD family that uses Fe or Mn as cofactor.

RESULTS: The SODa webtool http://babylone.ulb.ac.be/soda predicts if a target

sequence corresponds to an Fe/Mn SOD. If so, it predicts the metal ion

specificity (Fe, Mn or cambialistic) and the oligomerization mode (dimer or

tetramer) of the target. In addition, SODa proposes a list of residue

substitutions likely to improve the predicted preferences for the metal cofactor

and oligomerization mode. The method is based on residue fingerprints, consisting

of residues conserved in SOD sequences or typical of SOD subgroups, and of

interaction fingerprints, containing residue pairs that are in contact in SOD

structures.

CONCLUSION: SODa is shown to outperform and to be more discriminative than

traditional techniques based on pairwise sequence alignments. Moreover, the fact

that it proposes selected mutations makes it a valuable tool for rational protein

design.

DOI: 10.1186/1471-2105-9-257

PMCID: PMC2442099

PMID: 18518943 [Indexed for MEDLINE]

2467. Bioinformatics. 2008 Jun 1;24(11):1394-6. doi: 10.1093/bioinformatics/btn137.

Epub 2008 May 3.

Mireval: a web tool for simple microRNA prediction in genome sequences.

Ritchie W(1), Théodule FX, Gautheret D.

Author information:

(1)INSERM U928, Technologies Avancées pour le Génome et la Clinique, Luminy Case

906, 13288 Marseille, Cedex 09, France. W.Ritchie@centenary.org.au

SUMMARY: We have developed an online tool called mirEval which can search

sequences of up to 10 000 nt for novel microRNAs in multiple organisms. It is a

comprehensive tool, easy to use and very informative. It will allow users with no

prior knowledge of in-silico detection of microRNAs to take advantage of the most

successful approaches to investigate sequences of interest.

AVAILABILITY: The mirEval web server is available at

http://tagc.univ-mrs.fr/mireval

DOI: 10.1093/bioinformatics/btn137

PMID: 18453555 [Indexed for MEDLINE]

2468. Bioinformatics. 2008 Jun 1;24(11):1401-2. doi: 10.1093/bioinformatics/btn132.

Epub 2008 Apr 21.

OnD-CRF: predicting order and disorder in proteins using [corrected] conditional

random fields.

Wang L(1), Sauer UH.

Author information:

(1)Umeå Centre for Molecular Pathogenesis, UCMP, and Centre for Chemical Biology,

KBC, Umeå University, SE-901 87 Umeå, Sweden.

Erratum in

Bioinformatics. 2008 Jul 1;24(13):1562.

MOTIVATION: Order and Disorder prediction using Conditional Random Fields

(OnD-CRF) is a new method for accurately predicting the transition between

structured and mobile or disordered regions in proteins. OnD-CRF applies CRFs

relying on features which are generated from the amino acids sequence and from

secondary structure prediction. Benchmarking results based on CASP7 targets, and

evaluation with respect to several CASP criteria, rank the OnD-CRF model highest

among the fully automatic server group.

AVAILABILITY: http://babel.ucmp.umu.se/ond-crf/

DOI: 10.1093/bioinformatics/btn132

PMCID: PMC2387219

PMID: 18430742 [Indexed for MEDLINE]

2469. Comput Biol Chem. 2008 Jun;32(3):227-31. doi: 10.1016/j.compbiolchem.2008.03.002.

Epub 2008 Mar 18.

TMBETADISC-RBF: Discrimination of beta-barrel membrane proteins using RBF

networks and PSSM profiles.

Ou YY(1), Gromiha MM, Chen SA, Suwa M.

Author information:

(1)Department of Computer Science and Engineering, Yuan Ze University, Chung-Li,

Taiwan. yien@csie.org

Discriminating outer membrane proteins (OMPs) from other folding types of

globular and membrane proteins is an important task both for identifying OMPs

from genomic sequences and for the successful prediction of their secondary and

tertiary structures. We have developed a method based on radial basis function

networks and position specific scoring matrix (PSSM) profiles generated by

PSI-BLAST and non-redundant protein database. Our approach with PSSM profiles has

correctly predicted the OMPs with a cross-validated accuracy of 96.4% in a set of

1251 proteins, which contain 206 OMPs, 667 globular proteins and 378

alpha-helical inner membrane proteins. Furthermore, we applied our method on a

dataset containing 114 OMPs, 187 TMH proteins and 195 globular proteins obtained

with less than 20% sequence identity and obtained the cross-validated accuracy of

95%. This accuracy of discriminating OMPs is higher than other methods in the

literature and our method could be used as an effective tool for dissecting OMPs

from genomic sequences. We have developed a prediction server, TMBETADISC-RBF,

which is available at http://rbf.bioinfo.tw/~sachen/OMP.html.

DOI: 10.1016/j.compbiolchem.2008.03.002

PMID: 18434251 [Indexed for MEDLINE]

2470. J Comput Biol. 2008 Jun;15(5):505-24. doi: 10.1089/cmb.2008.0075.

Feedback algorithm and web-server for protein structure alignment.

Zhao Z(1), Fu B, Alanis FJ, Summa CM.

Author information:

(1)Department of Computer Science, University of New Orleans, New Orleans,

Louisiana, USA.

We have developed a feedback algorithm for protein structure alignment that uses

a series of phases to improve the global alignment between two protein backbones.

The method implements a self-improving learning strategy by sending the output of

one phase, the global alignment, to the next phase as an input. A web portal

implementing this method has been constructed and is freely available for use at

http://fpsa.cs.uno.edu/. Based on hundreds of test cases, we compare our

algorithm with three other, commonly used methods: CE, Dali, and SSM. Our results

show that, in most cases, our algorithm outputs a larger number of aligned

positions when the (C(alpha)) RMSD is comparable. Also, in many cases where the

number of aligned positions is larger or comparable to the other methods, our

learning method is able to achieve a smaller (C(alpha)) RMSD than the other

methods tested.

DOI: 10.1089/cmb.2008.0075

PMID: 18549304 [Indexed for MEDLINE]

2471. Biochem Biophys Res Commun. 2008 May 16;369(4):1166-8. doi:

10.1016/j.bbrc.2008.03.008. Epub 2008 Mar 13.

Prediction of translation initiation sites in human mRNA sequences with AUG start

codon in weak Kozak context: A neural network approach.

Tikole S(1), Sankararamakrishnan R.

Author information:

(1)Department of Biological Sciences and Bioengineering, Indian Institute of

Technology, Kanpur, Uttar Pradesh 208 016, India.

Translation of eukaryotic mRNAs is often regulated by nucleotides around the

start codon. A purine at position -3 and a guanine at position +4 contribute

significantly to enhance the translation efficiency. Algorithms to predict the

translation initiation site often fail to predict the start site if the sequence

context is not present. We have developed a neural network method to predict the

initiation site of mRNA sequences that lack the preferred nucleotides at the

positions -3 and +4 surrounding the translation initiation site. Neural networks

of various architectures comprising different number of hidden layers were

designed and tested for various sizes of windows of nucleotides surrounding

translation initiation sites. We found that the neural network with two hidden

layers showed a sensitivity of 83% and specificity of 73% indicating a vastly

improved performance in successfully predicting the translation initiation site

of mRNA sequences with weak Kozak context. WeakAUG server is freely available at

http://bioinfo.iitk.ac.in/AUGPred/.

DOI: 10.1016/j.bbrc.2008.03.008

PMID: 18342624 [Indexed for MEDLINE]

2472. Bioinformatics. 2008 May 15;24(10):1251-6. doi: 10.1093/bioinformatics/btn118.

Epub 2008 Apr 8.

Domain annotation of trimeric autotransporter adhesins--daTAA.

Szczesny P(1), Lupas A.

Author information:

(1)Department of Protein Evolution, Max-Planck Institute for Developmental

Biology, Spemannstr 35, 72076 Tuebingen, Germany.

MOTIVATION: Trimeric autotransporter adhesins (TAAs), such as Yersinia YadA,

Neisseria NadA, Moraxella UspAs, Haemophilus Hia and Bartonella BadA, are

important pathogenicity factors of proteobacteria. Their high sequence diversity

and distinct mosaic-like structure lead to difficulties in the annotation of

their sequences. These stem from the large number of short repeats, the presence

of compositionally unusual coiled-coils, fuzzy domain boundaries and regions of

seemingly low sequence complexity.

RESULTS: We have developed a workflow, named daTAA, for the accurate domain

annotation of TAAs. Its core consists of manually curated alignments and of

knowledge-based rules that enhance assignments made by sequence similarity.

Compared to general domain annotation servers such as PFAM, daTAA captures more

domains and provides more sensitive domain detection, as well as integrated and

detailed coiled-coil assignments.

AVAILABILITY: The daTAA server is freely accessible at

http://toolkit.tuebingen.mpg.de/dataa

DOI: 10.1093/bioinformatics/btn118

PMCID: PMC2373917

PMID: 18397894 [Indexed for MEDLINE]

2473. Bioinformatics. 2008 May 15;24(10):1316-7. doi: 10.1093/bioinformatics/btn121.

Epub 2008 Apr 7.

siRNA specificity searching incorporating mismatch tolerance data.

Chalk AM(1), Sonnhammer EL.

Author information:

(1)Department of Cell and Molecular Biology, Karolinska Institutet, S-171 77

Stockholm, Sweden.

Artificially synthesized short interfering RNAs (siRNAs) are widely used in

functional genomics to knock down specific target genes. One ongoing challenge is

to guarantee that the siRNA does not elicit off-target effects. Initial reports

suggested that siRNAs were highly sequence-specific; however, subsequent data

indicates that this is not necessarily the case. It is still uncertain what level

of similarity and other rules are required for an off-target effect to be

observed, and scoring schemes have not been developed to look beyond simple

measures such as the number of mismatches or the number of consecutive matching

bases present. We created design rules for predicting the likelihood of a

non-specific effect and present a web server that allows the user to check the

specificity of a given siRNA in a flexible manner using a combination of methods.

The server finds potential off-target matches in the corresponding RefSeq

database and ranks them according to a scoring system based on experimental

studies of specificity.AVAILABILITY: The server is available at

http://informatics-eskitis.griffith.edu.au/SpecificityServer.

DOI: 10.1093/bioinformatics/btn121

PMID: 18397893 [Indexed for MEDLINE]

2474. Bioinformatics. 2008 May 15;24(10):1271-7. doi: 10.1093/bioinformatics/btn114.

Epub 2008 Apr 3.

Prediction of the translocon-mediated membrane insertion free energies of protein

sequences.

Park Y(1), Helms V.

Author information:

(1)Center for Bioinformatics, Saarland University, Germany.

MOTIVATION: Helical membrane proteins (HMPs) play crucial roles in a variety of

cellular processes. Unlike water-soluble proteins, HMPs need not only to fold but

also get inserted into the membrane to be fully functional. This process of

membrane insertion is mediated by the translocon complex. Thus, it is of great

interest to develop computational methods for predicting the translocon-mediated

membrane insertion free energies of protein sequences.

RESULT: We have developed Membrane Insertion (MINS), a novel sequence-based

computational method for predicting the membrane insertion free energies of

protein sequences. A benchmark test gives a correlation coefficient of 0.74

between predicted and observed free energies for 357 known cases, which

corresponds to a mean unsigned error of 0.41 kcal/mol. These results are

significantly better than those obtained by traditional hydropathy analysis.

Moreover, the ability of MINS to reasonably predict membrane insertion free

energies of protein sequences allows for effective identification of

transmembrane (TM) segments. Subsequently, MINS was applied to predict the

membrane insertion free energies of 316 TM segments found in known structures. An

in-depth analysis of the predicted free energies reveals a number of interesting

findings about the biogenesis and structural stability of HMPs.

AVAILABILITY: A web server for MINS is available at

http://service.bioinformatik.uni-saarland.de/mins

DOI: 10.1093/bioinformatics/btn114

PMID: 18388143 [Indexed for MEDLINE]

2475. BMC Struct Biol. 2008 May 13;8:25. doi: 10.1186/1472-6807-8-25.

K2D2: estimation of protein secondary structure from circular dichroism spectra.

Perez-Iratxeta C(1), Andrade-Navarro MA.

Author information:

(1)Ontario Genomics Innovation Centre, Ottawa Health Research Institute, 501

Smyth, Ottawa, ON, K1H 8L6, Canada. cperez-iratxeta@ohri.ca

BACKGROUND: Circular dichroism spectroscopy is a widely used technique to analyze

the secondary structure of proteins in solution. Predictive methods use the

circular dichroism spectra from proteins of known tertiary structure to assess

the secondary structure contents of a protein with unknown structure given its

circular dichroism spectrum.

RESULTS: We developed K2D2, a method with an associated web server to estimate

protein secondary structure from circular dichroism spectra. The method uses a

self-organized map of spectra from proteins with known structure to deduce a map

of protein secondary structure that is used to do the predictions.

CONCLUSION: The K2D2 server is publicly accessible at

http://www.ogic.ca/projects/k2d2/. It accepts as input a circular dichroism

spectrum and outputs the estimated secondary structure content (alpha-helix and

beta-strand) of the corresponding protein, as well as an estimated measure of

error.

DOI: 10.1186/1472-6807-8-25

PMCID: PMC2397409

PMID: 18477405 [Indexed for MEDLINE]

2476. Bioinformatics. 2008 May 1;24(9):1145-53. doi: 10.1093/bioinformatics/btn097.

Epub 2008 Mar 12.

Optimizing the size of the sequence profiles to increase the accuracy of protein

sequence alignments generated by profile-profile algorithms.

Poleksic A(1), Fienup M.

Author information:

(1)Department of Computer Science, University of Northern Iowa, Cedar Falls, IA

50614, USA. poleksic@cs.uni.edu

MOTIVATION: Profile-based protein homology detection algorithms are valuable

tools in genome annotation and protein classification. By utilizing information

present in the sequences of homologous proteins, profile-based methods are often

able to detect extremely weak relationships between protein sequences, as

evidenced by the large-scale benchmarking experiments such as CASP and LiveBench.

RESULTS: We study the relationship between the sensitivity of a profile-profile

method and the size of the sequence profile, which is defined as the average

number of different residue types observed at the profile's positions. We also

demonstrate that improvements in the sensitivity of a profile-profile method can

be made by incorporating a profile-dependent scoring scheme, such as

position-specific background frequencies. The techniques presented in this

article are implemented in an alignment algorithm UNI-FOLD. When tested against

other well-established methods for fold recognition, UNI-FOLD shows increased

sensitivity and specificity in detecting remote relationships between protein

sequences.

AVAILABILITY: UNI-FOLD web server can be accessed at http://blackhawk.cs.uni.edu

DOI: 10.1093/bioinformatics/btn097

PMID: 18337259 [Indexed for MEDLINE]

2477. Bioinformatics. 2008 May 1;24(9):1137-44. doi: 10.1093/bioinformatics/btn093.

Epub 2008 Mar 7.

3D-Garden: a system for modelling protein-protein complexes based on

conformational refinement of ensembles generated with the marching cubes

algorithm.

Lesk VI(1), Sternberg MJ.

Author information:

(1)Division of Molecular Biosciences, Imperial College London, South Kensington,

London SW72AZ, UK. v.lesk@ic.ac.uk

MOTIVATION: Reliable structural modelling of protein-protein complexes has

widespread application, from drug design to advancing our knowledge of protein

interactions and function. This work addresses three important issues in

protein-protein docking: implementing backbone flexibility, incorporating prior

indications from experiment and bioinformatics, and providing public access via a

server. 3D-Garden (Global And Restrained Docking Exploration Nexus), our

benchmarked and server-ready flexible docking system, allows sophisticated

programming of surface patches by the user via a facet representation of the

interactors' molecular surfaces (generated with the marching cubes algorithm).

Flexibility is implemented as a weighted exhaustive conformer search for each

clashing pair of molecular branches in a set of 5000 models filtered from around

approximately 340,000 initially.

RESULTS: In a non-global assessment, carried out strictly according to the

protocols for number of models considered and model quality of the Critical

Assessment of Protein Interactions (CAPRI) experiment, over the widely-used

Benchmark 2.0 of 84 complexes, 3D-Garden identifies a set of ten models

containing an acceptable or better model in 29/45 test cases, including one with

large conformational change. In 19/45 cases an acceptable or better model is

ranked first or second out of 340,000 candidates.

AVAILABILITY: http://www.sbg.bio.ic.ac.uk/3dgarden (server).

DOI: 10.1093/bioinformatics/btn093

PMID: 18326508 [Indexed for MEDLINE]

2478. Brief Bioinform. 2008 May;9(3):243-9. doi: 10.1093/bib/bbm063. Epub 2008 Jan 16.

Two interactive Bioinformatics courses at the Bielefeld University Bioinformatics

Server.

Sczyrba A(1), Konermann S, Giegerich R.

Author information:

(1)Faculty of Technology, Bielefeld University, D-33594 Bielefeld, Germany.

asczyrba@TechFak.Uni-Bielefeld.DE

Conferences in computational biology continue to provide tutorials on classical

and new methods in the field. This can be taken as an indicator that education is

still a bottleneck in our field's process of becoming an established scientific

discipline. Bielefeld University has been one of the early providers of

bioinformatics education, both locally and via the internet. The Bielefeld

Bioinformatics Server (BiBiServ) offers a variety of older and new materials.

Here, we report on two online courses made available recently, one introductory

and one on the advanced level: (i) SADR: Sequence Analysis with Distributed

Resources (http://bibiserv.techfak.uni-bielefeld.de/sadr/) and (ii) ADP:

Algebraic Dynamic Programming in Bioinformatics

(http://bibiserv.techfak.uni-bielefeld.de/dpcourse/).

DOI: 10.1093/bib/bbm063

PMID: 18199576 [Indexed for MEDLINE]

2479. Comput Biol Med. 2008 May;38(5):620-2. doi: 10.1016/j.compbiomed.2008.02.009.

Epub 2008 Apr 9.

PubMedAlertMe--standalone Windows-based PubMed SDI software application.

Ma'ayan A(1).

Author information:

(1)Department of Pharmacology and Systems Therapeutics, Mount Sinai School of

Medicine, New York, NY 10029, USA. avi.maayan@mssm.edu <avi.maayan@mssm.edu>

PubMedAlertMe is a Windows-based software system for automatically receiving

e-mail alert messages about recent publications listed on PubMed. The e-mail

messages contain links to newly available abstracts listed on PubMed describing

publications that were selectively returned from a specified list of queries.

Links are also provided to directly export citations to EndNote, and links are

provided to directly forward articles to colleagues. The program is standalone.

Thus, it does not require a remote mail server or user registration.

PubMedAlertMe is free software, and can be downloaded from:

http://amp.pharm.mssm.edu/PubMedAlertMe/PubMedAlertMe\_setup.zip.

DOI: 10.1016/j.compbiomed.2008.02.009

PMCID: PMC2431148

PMID: 18402930 [Indexed for MEDLINE]

2480. Comput Methods Programs Biomed. 2008 May;90(2):148-53. doi:

10.1016/j.cmpb.2007.12.003. Epub 2008 Feb 7.

RISP: a web-based server for prediction of RNA-binding sites in proteins.

Tong J(1), Jiang P, Lu ZH.

Author information:

(1)Department of Biological Science and Medical Engineering, Southeast

University, Nanjing 210096, China. zhlu@seu.edu.cn

Protein-RNA interactions play significant roles in a number of biological

activities, such as protein synthesis, regulation of gene expression. Here we

propose a hybrid RISP (RNA-interaction site prediction) method, using support

vector machine (SVM) in conjunction with evolutionary information of amino acid

sequences in terms of their position-specific scoring matrices (PSSMs) for

prediction of RNA-binding sites. The results show that our RISP method has 72.2%

net prediction (NP) (61.0% sensitivity and 83.3% specificity). When compared with

previous studies, this novel method appears more accurate and better

generalization abilities. RISP is freely available at http://grc.seu.edu.cn/RISP.

Given a protein sequence, RISP decides whether residue in the protein is

RNA-binding or not (optimal prediction), and gives the confidence value, 'high

specificity' prediction and 'high sensitivity' prediction.

DOI: 10.1016/j.cmpb.2007.12.003

PMID: 18261823 [Indexed for MEDLINE]

2481. Fungal Genet Biol. 2008 May;45(5):638-45. doi: 10.1016/j.fgb.2008.01.004. Epub

2008 Jan 26.

FOLy: an integrated database for the classification and functional annotation of

fungal oxidoreductases potentially involved in the degradation of lignin and

related aromatic compounds.

Levasseur A(1), Piumi F, Coutinho PM, Rancurel C, Asther M, Delattre M, Henrissat

B, Pontarotti P, Asther M, Record E.

Author information:

(1)UMR 1163 INRA de Biotechnologie des Champignons Filamenteux, IFR86-BAIM,

Universités de Provence et de la Méditerranée, ESIL, 163 Avenue de Luminy, Case

Postale 925, 13288 Marseille Cedex 09, France.

The breakdown of lignin by fungi is a key step during carbon recycling in

terrestrial ecosystems. This process is of great interest for green and white

biotechnological applications. Given the importance of these enzymatic processes,

we have classified the enzymes potentially involved in lignin catabolism into

sequence-based families and integrated them in a newly developed database,

designated Fungal Oxidative Lignin enzymes (FOLy). Families were defined after

sequence similarity searches starting from protein sequences and validated by the

convergence of results with biochemical experiments reported in the literature.

The resulting database was applied as a tool for the functional annotation of

genomes from different fungi, namely (i) the Basidiomycota Coprinopsis cinerea,

Phanerochaete chrysosporium and Ustilago maydis and (ii) the Ascomycota

Aspergillus nidulans and Trichoderma reesei. Genomic comparison of the

oxidoreductases of these fungi revealed significant differences in the putative

enzyme arsenals. Two Ascomycota fungal genomes were annotated and new candidate

genes were identified that could be useful for lignin degradation and (or)

melanin synthesis, and their function investigated experimentally. This database

efforts aims at providing the means to get new insights for the understanding and

biotechnological exploitation of the lignin degradation. A WWW server giving

access to the routinely updated FOLy classifications of enzymes potentially

involved in lignin degradation can be found at http://foly.esil.univ-mrs.fr.

DOI: 10.1016/j.fgb.2008.01.004

PMID: 18308593 [Indexed for MEDLINE]

2482. Proteins. 2008 May 1;71(2):920-37.

Protein structure mining using a structural alphabet.

Tyagi M(1), de Brevern AG, Srinivasan N, Offmann B.

Author information:

(1)Laboratoire de Biochimie et Génétique Moléculaire, Bioinformatics Team,

Université de La Réunion, BP 7151, 15 avenue René Cassin, 97715 Saint Denis

Messag Cedex 09, La Réunion, France.

We present a comprehensive evaluation of a new structure mining method called

PB-ALIGN. It is based on the encoding of protein structure as 1D sequence of a

combination of 16 short structural motifs or protein blocks (PBs). PBs are short

motifs capable of representing most of the local structural features of a protein

backbone. Using derived PB substitution matrix and simple dynamic programming

algorithm, PB sequences are aligned the same way amino acid sequences to yield

structure alignment. PBs are short motifs capable of representing most of the

local structural features of a protein backbone. Alignment of these local

features as sequence of symbols enables fast detection of structural similarities

between two proteins. Ability of the method to characterize and align regions

beyond regular secondary structures, for example, N and C caps of helix and loops

connecting regular structures, puts it a step ahead of existing methods, which

strongly rely on secondary structure elements. PB-ALIGN achieved efficiency of

85% in extracting true fold from a large database of 7259 SCOP domains and was

successful in 82% cases to identify true super-family members. On comparison to

13 existing structure comparison/mining methods, PB-ALIGN emerged as the best on

general ability test dataset and was at par with methods like YAKUSA and CE on

nontrivial test dataset. Furthermore, the proposed method performed well when

compared to flexible structure alignment method like FATCAT and outperforms in

processing speed (less than 45 s per database scan). This work also establishes a

reliable cut-off value for the demarcation of similar folds. It finally shows

that global alignment scores of unrelated structures using PBs follow an extreme

value distribution. PB-ALIGN is freely available on web server called Protein

Block Expert (PBE) at http://bioinformatics.univ-reunion.fr/PBE/.

DOI: 10.1002/prot.21776

PMID: 18004784 [Indexed for MEDLINE]

2483. Anal Biochem. 2008 Apr 15;375(2):388-90. doi: 10.1016/j.ab.2008.01.012. Epub 2008

Jan 15.

HIVcleave: a web-server for predicting human immunodeficiency virus protease

cleavage sites in proteins.

Shen HB(1), Chou KC.

Author information:

(1)Gordon Life Science Institute, San Diego, CA 92130, USA.

According to the ''distorted key theory'' [K.C. Chou, Analytical Biochemistry,

233 (1996) 1-14], the information of cleavage sites of proteins by HIV (human

immunodeficiency virus) protease is very useful for finding effective inhibitors

against HIV, the culprit of AIDS (acquired immunodeficiency syndrome). To meet

the increasing need in this regard, a web-server called HIVcleave was established

at http://chou.med.harvard.edu/bioinf/HIV/. In this note we provide a

step-to-step guide for how to use HIVcleave to identify the cleavage sites of a

query protein sequence by HIV-1 and HIV-2 proteases, respectively.

DOI: 10.1016/j.ab.2008.01.012

PMID: 18249180 [Indexed for MEDLINE]

2484. Bioinformatics. 2008 Apr 15;24(8):1115-7. doi: 10.1093/bioinformatics/btn086.

Epub 2008 Mar 5.

e-LiSe--an online tool for finding needles in the '(Medline) haystack'.

Gladki A(1), Siedlecki P, Kaczanowski S, Zielenkiewicz P.

Author information:

(1)Bioinformatics Department, Institute of Biochemistry and Biophysics, Polish

Academy of Sciences, ul. Pawinskiego 5a, 02-106, Warszawa, Poland.

Using literature databases one can find not only known and true relations between

processes but also less studied, non-obvious associations. The main problem with

discovering such type of relevant biological information is 'selection'. The

ability to distinguish between a true correlation (e.g. between different types

of biological processes) and random chance that this correlation is statistically

significant is crucial for any bio-medical research, literature mining being no

exception. This problem is especially visible when searching for information

which has not been studied and described in many publications. Therefore, a novel

bio-linguistic statistical method is required, capable of 'selecting' true

correlations, even when they are low-frequency associations. In this article, we

present such statistical approach based on Z-score and implemented in a web-based

application 'e-LiSe'.AVAILABILITY: The software is available at

http://miron.ibb.waw.pl/elise/

DOI: 10.1093/bioinformatics/btn086

PMID: 18321884 [Indexed for MEDLINE]

2485. Biochimie. 2008 Apr;90(4):563-9. Epub 2007 Sep 29.

DigiPINS: a database for vertebrate exonic single nucleotide polymorphisms and

its application to cancer association studies.

Navratil V(1), Penel S, Delmotte S, Mouchiroud D, Gautier C, Aouacheria A.

Author information:

(1)Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard Lyon

1, 69622 Villeurbanne Cedex, France. navratil@biomserve.univ-lyon1.fr

Single nucleotide polymorphisms (SNPs), which are the most abundant form of

genetic variations in numerous organisms, have emerged as important tools for the

study of complex genetic traits and deciphering of genome evolution.

High-throughput genome sequencing projects worldwide provide an unprecedented

opportunity for whole-genome SNP analysis in a variety of species. To facilitate

SNP discovery in vertebrates, we have developed a web-based, user-friendly, and

fully automated application, DigiPINS, for genome-wide identification of exonic

SNPs from EST data. Currently, the database can be used to the mining of exonic

SNPs in six complete genomes (Homo sapiens, Mus musculus, Rattus norvegicus,

Canis familiaris, Gallus gallus and Danio rerio). In addition to providing

information on sequence conservation, DigiPINS allows compilation of

comprehensive sets of polymorphisms within cancer candidate genes or

identification of novel cancer markers, making it potentially useful for cancer

association studies. The DigiPINS server is available via the internet at

http://pbil.univ-lyon1.fr/gem/DigiPINS/query\_DigiPINS.php.

DOI: 10.1016/j.biochi.2007.09.017

PMID: 17988782 [Indexed for MEDLINE]

2486. Diabetologia. 2008 Apr;51(4):546-53. doi: 10.1007/s00125-008-0942-y. Epub 2008

Feb 23.

Best practice guidelines for the molecular genetic diagnosis of maturity-onset

diabetes of the young.

Ellard S(1), Bellanné-Chantelot C, Hattersley AT; European Molecular Genetics

Quality Network (EMQN) MODY group.

Collaborators: Carette C, Castano Gonzalez L, de Nanclares Leal G, Elles R,

Gaspar G, Gasperikova D, Hansen T, Herr M, Kamarainen O, Kannengiesser C, Klimes

I, Lacape G, Losekoot M, Malecki M, Meyer P, Njolstad P, Predragovic T, Pruhova

S, Wuyts W.

Author information:

(1)Institute of Biomedical and Clinical Science, Peninsula Medical School,

Exeter, UK. Sian.Ellard@rdeft.nhs.uk

AIMS/HYPOTHESIS: Mutations in the GCK and HNF1A genes are the most common cause

of the monogenic forms of diabetes known as 'maturity-onset diabetes of the

young'. GCK encodes the glucokinase enzyme, which acts as the pancreatic glucose

sensor, and mutations result in stable, mild fasting hyperglycaemia. A

progressive insulin secretory defect is seen in patients with mutations in the

HNF1A and HNF4A genes encoding the transcription factors hepatocyte nuclear

factor-1 alpha and -4 alpha. A molecular genetic diagnosis often changes

management, since patients with GCK mutations rarely require pharmacological

treatment and HNF1A/4A mutation carriers are sensitive to sulfonylureas. These

monogenic forms of diabetes are often misdiagnosed as type 1 or 2 diabetes. Best

practice guidelines for genetic testing were developed to guide testing and

reporting of results.

METHODS: A workshop was held to discuss clinical criteria for testing and the

interpretation of molecular genetic test results. The participants included 22

clinicians and scientists from 13 countries. Draft best practice guidelines were

formulated and edited using an online tool (http://www.coventi.com).

RESULTS: An agreed set of clinical criteria were defined for the testing of

babies, children and adults for GCK, HNF1A and HNF4A mutations. Reporting

scenarios were discussed and consensus statements produced.

CONCLUSIONS/INTERPRETATION: Best practice guidelines have been established for

monogenic forms of diabetes caused by mutations in the GCK, HNF1A and HNF4A

genes. The guidelines include both diagnostic and predictive genetic tests and

interpretation of the results.

DOI: 10.1007/s00125-008-0942-y

PMCID: PMC2270360

PMID: 18297260 [Indexed for MEDLINE]

2487. J Chem Inf Model. 2008 Apr;48(4):691-703. doi: 10.1021/ci700334f. Epub 2008 Apr

11.

Distributed chemical computing using ChemStar: an open source java remote method

invocation architecture applied to large scale molecular data from PubChem.

Karthikeyan M(1), Krishnan S, Pandey AK, Bender A, Tropsha A.

Author information:

(1)Digital Information Resource Center, Information Division, National Chemical

Laboratory, Pune 411008, India. m.karthikeyan@ncl.res.in

We present the application of a Java remote method invocation (RMI) based open

source architecture to distributed chemical computing. This architecture was

previously employed for distributed data harvesting of chemical information from

the Internet via the Google application programming interface (API; ChemXtreme).

Due to its open source character and its flexibility, the underlying

server/client framework can be quickly adopted to virtually every computational

task that can be parallelized. Here, we present the server/client communication

framework as well as an application to distributed computing of chemical

properties on a large scale (currently the size of PubChem; about 18 million

compounds), using both the Marvin toolkit as well as the open source JOELib

package. As an application, for this set of compounds, the agreement of log P and

TPSA between the packages was compared. Outliers were found to be mostly

non-druglike compounds and differences could usually be explained by differences

in the underlying algorithms. ChemStar is the first open source distributed

chemical computing environment built on Java RMI, which is also easily adaptable

to user demands due to its "plug-in architecture". The complete source codes as

well as calculated properties along with links to PubChem resources are available

on the Internet via a graphical user interface at

http://moltable.ncl.res.in/chemstar/.

DOI: 10.1021/ci700334f

PMID: 18402434

2488. J Undergrad Neurosci Educ. 2008 Spring;6(2):A68-73. Epub 2008 Jun 15.

Online multimedia teaching tool for Parkinson's disease.

Misiaszek G(1), Riconscente M, Henke M, Walsh JP.

Author information:

(1)Davis School of Gerontology.

We developed an online multimedia tool designed to enhance the student-learning

environment in neurosciences through multi-sensory engagement. The combined use

of scrolling text, narrations, and visual imagery engages multiple sensory

modalities for effective learning, and it assists students in visualizing complex

processes in the nervous system. The initial rollout of the online tool is for

instruction in Parkinson's disease (PD), but its structure is flexible and can be

used for teaching a variety of subjects. The instructor can access the tool

online during lecture, and students can access the same information via the

internet outside of class. In addition, each chapter is stand-alone and thus can

be accessed online by other faculty or students to supplement other courses.

Within each chapter or module, information is presented in outline format with

greater detail accessible via sequential drop-down menus. This layering of

related topics creates a spatial and motor-accessed path for learning. These

multiple forms of engagement offer rich information representations to improve

students' knowledge encoding, storing, and retrieval via multiple pathways. For

instance, the tool includes student-controlled 2-D and 3-D animations, and video

clip demonstrations of both patient case studies and on-campus research projects

directly related to the subject material. Supplemental readings consist of

current research articles (in Adobe Acrobat PDF file format) accessed within each

educational topic. The teaching tool for PD is online at

http://geroauen.usc.edu/Gero414\_Beta/.

PMCID: PMC3592663

PMID: 23493487

2489. Protein Eng Des Sel. 2008 Apr;21(4):279-82. doi: 10.1093/protein/gzn006. Epub

2008 Feb 20.

CytoPred: a server for prediction and classification of cytokines.

Lata S(1), Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India.

Cytokines are messengers of immune system. They are small secreted proteins that

mediate and regulate the immune system, inflammation and hematopoiesis. Recent

studies have revealed important roles played by the cytokines in adjuvants as

therapeutic targets and in cancer therapy. In this paper, an attempt has been

made to predict this important class of proteins and classify further them into

families and subfamilies. A PSI-BLAST+Support Vector Machine-based hybrid

approach is adopted to develop the prediction methods. CytoPred is capable of

predicting cytokines with an accuracy of 98.29%. The overall accuracy of

classification of cytokines into four families and further classification into

seven subfamilies is 99.77 and 97.24%, respectively. It has been shown by

comparison that CytoPred performs better than the already existing CTKPred. A

user-friendly server CytoPred has been developed and available at

http://www.imtech.res.in/raghava/cytopred.

DOI: 10.1093/protein/gzn006

PMID: 18287174 [Indexed for MEDLINE]

2490. Proteins. 2008 Apr;71(1):61-7.

Assessing secondary structure assignment of protein structures by using pairwise

sequence-alignment benchmarks.

Zhang W(1), Dunker AK, Zhou Y.

Author information:

(1)Indiana University School of Informatics, Indiana University-Purdue

University, Indianapolis, Indiana 46202, USA.

How to make an objective assignment of secondary structures based on a protein

structure is an unsolved problem. Defining the boundaries between helix, sheet,

and coil structures is arbitrary, and commonly accepted standard assignments do

not exist. Here, we propose a criterion that assesses secondary structure

assignment based on the similarity of the secondary structures assigned to

pairwise sequence-alignment benchmarks, where these benchmarks are determined by

prior structural alignments of the protein pairs. This criterion is used to rank

six secondary structure assignment methods: STRIDE, DSSP, SECSTR, KAKSI, P-SEA,

and SEGNO with three established sequence-alignment benchmarks (PREFAB, SABmark,

and SALIGN). STRIDE and KAKSI achieve comparable success rates in assigning the

same secondary structure elements to structurally aligned residues in the three

benchmarks. Their success rates are between 1-4% higher than those of the other

four methods. The consensus of STRIDE, KAKSI, SECSTR, and P-SEA, called SKSP,

improves assignments over the best single method in each benchmark by an

additional 1%. These results support the usefulness of the sequence-alignment

benchmarks as a means to evaluate secondary structure assignment. The SKSP server

and the benchmarks can be accessed at http://sparks.informatics.iupui.edu

(c) 2007 Wiley-Liss, Inc.

DOI: 10.1002/prot.21654

PMID: 17932927 [Indexed for MEDLINE]

2491. Proteins. 2008 Apr;71(1):189-94.

Prediction of RNA binding sites in a protein using SVM and PSSM profile.

Kumar M(1), Gromiha MM, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Chandigarh-160036,

India.

RNA-binding proteins (RBPs) play key roles in post-transcriptional control of

gene expression, which, along with transcriptional regulation, is a major way to

regulate patterns of gene expression during development. Thus, the identification

and prediction of RNA binding sites is an important step in comprehensive

understanding of how RBPs control organism development. Combining evolutionary

information and support vector machine (SVM), we have developed an improved

method for predicting RNA binding sites or RNA interacting residues in a protein

sequence. The prediction models developed in this study have been trained and

tested on 86 RNA binding protein chains and evaluated using fivefold cross

validation technique. First, a SVM model was developed that achieved a maximum

Matthew's correlation coefficient (MCC) of 0.31. The performance of this SVM

model further improved the MCC from 0.31 to 0.45, when multiple sequence

alignment in the form of PSSM profiles was used as input to the SVM, which is far

better than the maximum MCC achieved by previous methods (0.41) on the same

dataset. In addition, SVM models were also developed on an alternative dataset

that contained 107 RBP chains. Utilizing PSSM as input information to the SVM,

the training/testing on this alternate dataset achieved a maximum MCC of 0.32.

Conclusively, the prediction performance of SVM models developed in this study is

better than the existing methods on the same datasets. A web server 'Pprint' was

also developed for predicting RNA binding residues in a protein sequence which is

freely available at http://www.imtech.res.in/raghava/pprint/.

(c) 2007 Wiley-Liss, Inc.

DOI: 10.1002/prot.21677

PMID: 17932917 [Indexed for MEDLINE]

2492. BMC Syst Biol. 2008 Mar 29;2:30. doi: 10.1186/1752-0509-2-30.

Optimal enumeration of state space of finitely buffered stochastic molecular

networks and exact computation of steady state landscape probability.

Cao Y(1), Liang J.

Author information:

(1)Shanghai Center for Systems Biomedicine (SCSB), Shanghai Jiao Tong University,

Shanghai, 200240, China. yfcao@sjtu.edu.cn

BACKGROUND: Stochasticity plays important roles in many molecular networks when

molecular concentrations are in the range of 0.1 muM to 10nM (about 100 to 10

copies in a cell). The chemical master equation provides a fundamental framework

for studying these networks, and the time-varying landscape probability

distribution over the full microstates, i.e., the combination of copy numbers of

molecular species, provide a full characterization of the network dynamics. A

complete characterization of the space of the microstates is a prerequisite for

obtaining the full landscape probability distribution of a network. However,

there are neither closed-form solutions nor algorithms fully describing all

microstates for a given molecular network.

RESULTS: We have developed an algorithm that can exhaustively enumerate the

microstates of a molecular network of small copy numbers under the condition that

the net gain in newly synthesized molecules is smaller than a predefined limit.

We also describe a simple method for computing the exact mean or steady state

landscape probability distribution over microstates. We show how the full

landscape probability for the gene networks of the self-regulating gene and the

toggle-switch in the steady state can be fully characterized. We also give an

example using the MAPK cascade network. Data and server will be available at URL:

http://scsb.sjtu.edu.cn/statespace.

CONCLUSION: Our algorithm works for networks of small copy numbers buffered with

a finite copy number of net molecules that can be synthesized, regardless of the

reaction stoichiometry, and is optimal in both storage and time complexity. The

algorithm can also be used to calculate the rates of all transitions between

microstates from given reactions and reaction rates. The buffer size is limited

by the available memory or disk storage. Our algorithm is applicable to a class

of biological networks when the copy numbers of molecules are small and the

network is closed, or the network is open but the net gain in newly synthesized

molecules does not exceed a predefined buffer capacity. For these networks, our

method allows full stochastic characterization of the mean landscape probability

distribution, and the steady state when it exists.

DOI: 10.1186/1752-0509-2-30

PMCID: PMC2375859

PMID: 18373871 [Indexed for MEDLINE]

2493. BMC Bioinformatics. 2008 Mar 26;9:170. doi: 10.1186/1471-2105-9-170.

PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic

genomes.

Fong C(1), Rohmer L, Radey M, Wasnick M, Brittnacher MJ.

Author information:

(1)Department of Genome Sciences, University of Washington, Box 357710, Seattle,

Washington 98195, USA. cfong@u.washington.edu

BACKGROUND: The conservation of gene order among prokaryotic genomes can provide

valuable insight into gene function, protein interactions, or events by which

genomes have evolved. Although some tools are available for visualizing and

comparing the order of genes between genomes of study, few support an efficient

and organized analysis between large numbers of genomes. The Prokaryotic Sequence

homology Analysis Tool (PSAT) is a web tool for comparing gene neighborhoods

among multiple prokaryotic genomes.

RESULTS: PSAT utilizes a database that is preloaded with gene annotation, BLAST

hit results, and gene-clustering scores designed to help identify regions of

conserved gene order. Researchers use the PSAT web interface to find a gene of

interest in a reference genome and efficiently retrieve the sequence homologs

found in other bacterial genomes. The tool generates a graphic of the genomic

neighborhood surrounding the selected gene and the corresponding regions for its

homologs in each comparison genome. Homologs in each region are color coded to

assist users with analyzing gene order among various genomes. In contrast to

common comparative analysis methods that filter sequence homolog data based on

alignment score cutoffs, PSAT leverages gene context information for homologs,

including those with weak alignment scores, enabling a more sensitive analysis.

Features for constraining or ordering results are designed to help researchers

browse results from large numbers of comparison genomes in an organized manner.

PSAT has been demonstrated to be useful for helping to identify gene orthologs

and potential functional gene clusters, and detecting genome modifications that

may result in loss of function.

CONCLUSION: PSAT allows researchers to investigate the order of genes within

local genomic neighborhoods of multiple genomes. A PSAT web server for public use

is available for performing analyses on a growing set of reference genomes

through any web browser with no client side software setup or installation

required. Source code is freely available to researchers interested in setting up

a local version of PSAT for analysis of genomes not available through the public

server. Access to the public web server and instructions for obtaining source

code can be found at http://www.nwrce.org/psat.

DOI: 10.1186/1471-2105-9-170

PMCID: PMC2358893

PMID: 18366802 [Indexed for MEDLINE]

2494. BMC Bioinformatics. 2008 Mar 18;9:151. doi: 10.1186/1471-2105-9-151.

H2r: identification of evolutionary important residues by means of an entropy

based analysis of multiple sequence alignments.

Merkl R(1), Zwick M.

Author information:

(1)Institut für Biophysik und Physikalische Biochemie, Universität Regensburg,

D-93040 Regensburg, Germany. Rainer.Merkl@biologie.uni-regensburg.de

BACKGROUND: A multiple sequence alignment (MSA) generated for a protein can be

used to characterise residues by means of a statistical analysis of single

columns. In addition to the examination of individual positions, the

investigation of co-variation of amino acid frequencies offers insights into

function and evolution of the protein and residues.

RESULTS: We introduce conn(k), a novel parameter for the characterisation of

individual residues. For each residue k, conn(k) is the number of most extreme

signals of co-evolution. These signals were deduced from a normalised mutual

information (MI) value U(k, l) computed for all pairs of residues k, l. We

demonstrate that conn(k) is a more robust indicator than an individual MI-value

for the prediction of residues most plausibly important for the evolution of a

protein. This proposition was inferred by means of statistical methods. It was

further confirmed by the analysis of several proteins. A server, which computes

conn(k)-values is available at http://www-bioinf.uni-regensburg.de.

CONCLUSION: The algorithms H2r, which analyses MSAs and computes conn(k)-values,

characterises a specific class of residues. In contrast to strictly conserved

ones, these residues possess some flexibility in the composition of side chains.

However, their allocation is sensibly balanced with several other positions, as

indicated by conn(k).

DOI: 10.1186/1471-2105-9-151

PMCID: PMC2323388

PMID: 18366663 [Indexed for MEDLINE]

2495. Vet Microbiol. 2008 Mar 18;127(3-4):369-78. Epub 2007 Sep 22.

Design and implementation of a database for Brucella melitensis genome

annotation.

De Hertogh B(1), Lahlimi L, Lambert C, Letesson JJ, Depiereux E.

Author information:

(1)U.R.B.M., F.U.N.D.P., Rue de Bruxelles, 61, B-5000 Namur, Belgium.

The genome sequences of three Brucella biovars and of some species close to

Brucella sp. have become available, leading to new relationship analysis.

Moreover, the automatic genome annotation of the pathogenic bacteria Brucella

melitensis has been manually corrected by a consortium of experts, leading to 899

modifications of start sites predictions among the 3198 open reading frames

(ORFs) examined. This new annotation, coupled with the results of automatic

annotation tools of the complete genome sequences of the B. melitensis genome

(including BLASTs to 9 genomes close to Brucella), provides numerous data sets

related to predicted functions, biochemical properties and phylogenic

comparisons. To made these results available, alphaPAGe, a functional

auto-updatable database of the corrected sequence genome of B. melitensis, has

been built, using the entity-relationship (ER) approach and a multi-purpose

database structure. A friendly graphical user interface has been designed, and

users can carry out different kinds of information by three levels of queries:

(1) the basic search use the classical keywords or sequence identifiers; (2) the

original advanced search engine allows to combine (by using logical operators)

numerous criteria: (a) keywords (textual comparison) related to the pCDS's

function, family domains and cellular localization; (b) physico-chemical

characteristics (numerical comparison) such as isoelectric point or molecular

weight and structural criteria such as the nucleic length or the number of

transmembrane helix (TMH); (c) similarity scores with Escherichia coli and 10

species phylogenetically close to B. melitensis; (3) complex queries can be

performed by using a SQL field, which allows all queries respecting the

database's structure. The database is publicly available through a Web server at

the following url: http://www.fundp.ac.be/urbm/bioinfo/aPAGe.

DOI: 10.1016/j.vetmic.2007.09.010

PMID: 18160234 [Indexed for MEDLINE]

2496. Bioinformatics. 2008 Mar 15;24(6):807-14. doi: 10.1093/bioinformatics/btn039.

Epub 2008 Feb 1.

De novo identification of highly diverged protein repeats by probabilistic

consistency.

Biegert A(1), Söding J.

Author information:

(1)Department for Protein Evolution, Max Planck Institute for Developmental

Biology, Spemannstr. 35, 72076 Tübingen, Germany.

MOTIVATION: An estimated 25% of all eukaryotic proteins contain repeats, which

underlines the importance of duplication for evolving new protein functions.

Internal repeats often correspond to structural or functional units in proteins.

Methods capable of identifying diverged repeated segments or domains at the

sequence level can therefore assist in predicting domain structures, inferring

hypotheses about function and mechanism, and investigating the evolution of

proteins from smaller fragments.

RESULTS: We present HHrepID, a method for the de novo identification of repeats

in protein sequences. It is able to detect the sequence signature of structural

repeats in many proteins that have not yet been known to possess internal

sequence symmetry, such as outer membrane beta-barrels. HHrepID uses HMM-HMM

comparison to exploit evolutionary information in the form of multiple sequence

alignments of homologs. In contrast to a previous method, the new method (1)

generates a multiple alignment of repeats; (2) utilizes the transitive nature of

homology through a novel merging procedure with fully probabilistic treatment of

alignments; (3) improves alignment quality through an algorithm that maximizes

the expected accuracy; (4) is able to identify different kinds of repeats within

complex architectures by a probabilistic domain boundary detection method and (5)

improves sensitivity through a new approach to assess statistical significance.

AVAILABILITY: Server: http://toolkit.tuebingen.mpg.de/hhrepid; Executables:

ftp://ftp.tuebingen.mpg.de/pub/protevo/HHrepID

DOI: 10.1093/bioinformatics/btn039

PMID: 18245125 [Indexed for MEDLINE]

2497. Bioinformatics. 2008 Mar 15;24(6):863-5. doi: 10.1093/bioinformatics/btn043. Epub

2008 Jan 30.

Prophinder: a computational tool for prophage prediction in prokaryotic genomes.

Lima-Mendez G(1), Van Helden J, Toussaint A, Leplae R.

Author information:

(1)Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université

Libre de Bruxelles, 1050 Bruxelles, Belgium. gipsi@scmbb.ulb.ac.be

Prophinder is a prophage prediction tool coupled with a prediction database, a

web server and web service. Predicted prophages will help to fill the gaps in the

current sparse phage sequence space, which should cover an estimated 100 million

species. Systematic and reliable predictions will enable further studies of

prophages contribution to the bacteriophage gene pool and to better understand

gene shuffling between prophages and phages infecting the same host.AVAILABILITY:

Softare is available at http://aclame.ulb.ac.be/prophinder

DOI: 10.1093/bioinformatics/btn043

PMID: 18238785 [Indexed for MEDLINE]

2498. Bioinformatics. 2008 Mar 1;24(5):606-12. doi: 10.1093/bioinformatics/btn005. Epub

2008 Jan 9.

Reconstructing ancestral genome content based on symmetrical best alignments and

Dollo parsimony.

Sakarya O(1), Kosik KS, Oakley TH.

Author information:

(1)Department of Computer Science, University of California, Santa Barbara, CA

93106, USA.

MOTIVATION: Gene duplications and losses (GDLs) are important events in genome

evolution. They result in expansion or contraction of gene families, with a

likely role in phenotypic evolution. As more genomes become available and their

annotations are improved, software programs capable of rapidly and accurately

identifying the content of ancestral genomes and the timings of GDLs become

necessary to understand the unique evolution of each lineage.

RESULTS: We report EvolMAP, a new algorithm and software that utilizes a species

tree-based gene clustering method to join all-to-all symmetrical similarity

comparisons of multiple gene sets in order to infer the gene composition of

multiple ancestral genomes. The algorithm further uses Dollo parsimony-based

comparison of the inferred ancestral genes to pinpoint the timings of GDLs onto

evolutionary intervals marked by speciation events. Using EvolMAP, first we

analyzed the expansion of four families of G-protein coupled receptors (GPCRs)

within animal lineages. Additional to demonstrating the unique expansion tree for

each family, results also show that the ancestral eumetazoan genome contained

many fewer GPCRs than modern animals, and these families expanded through

concurrent lineage-specific duplications. Second, we analyzed the history of GDLs

in mammalian genomes by comparing seven proteomes. In agreement with previous

studies, we report that the mammalian gene family sizes have changed drastically

through their evolution. Interestingly, although we identified a potential source

of duplication for 75% of the gained genes, remaining 25% did not have clear-cut

sources, revealing thousands of genes that have likely gained their distinct

sequence identities within the descent of mammals.

AVAILABILITY: Query server, source code and executable are available at

http://kosik-web.mcdb.ucsb.edu/evolmap/index.htm .

DOI: 10.1093/bioinformatics/btn005

PMID: 18184685 [Indexed for MEDLINE]

2499. Bioinformatics. 2008 Mar 1;24(5):645-51. doi: 10.1093/bioinformatics/btm641. Epub

2008 Jan 5.

Structural search and retrieval using a tableau representation of protein folding

patterns.

Konagurthu AS(1), Stuckey PJ, Lesk AM.

Author information:

(1)Department of Biochemistry and Molecular Biology and The Huck Institute for

Genomics, Proteomics and Bioinformatics, The Pennsylvania State University,

University Park, PA 16802, USA. arun@bx.psu.edu

Comparison and classification of folding patterns from a database of protein

structures is crucial to understand the principles of protein architecture,

evolution and function. Current search methods for proteins with similar folding

patterns are slow and computationally intensive. The sharp growth in the number

of known protein structures poses severe challenges for methods of structural

comparison. There is a need for methods that can search the database of

structures accurately and rapidly. We provide several methods to search for

similar folding patterns using a concise tableau representation of proteins that

encodes the relative geometry of secondary structural elements. Our first

approach allows the extraction of identical and very closely-related protein

folding patterns in constant-time (per hit). Next, we address the hard

computational problem of extraction of maximally-similar subtableaux, when

comparing two tableaux. We solve the problem using Quadratic and Linear integer

programming formulations and demonstrate their power to identify subtle

structural similarities, especially when protein structures significantly

diverge. Finally, we describe a rapid and accurate method for comparing a query

structure against a database of protein domains, TableauSearch. TableauSearch is

rapid enough to search the entire structural database in seconds on a standard

desktop computer. Our analysis of TableauSearch on many queries shows that the

method is very accurate in identifying similarities of folding patterns, even

between distantly related proteins.AVAILABILITY: A web server implementing the

TableauSearch is available from http://hollywood.bx.psu.edu/TabSearch.

DOI: 10.1093/bioinformatics/btm641

PMID: 18175768 [Indexed for MEDLINE]

2500. Bioinformatics. 2008 Mar 1;24(5):613-20. doi: 10.1093/bioinformatics/btm626. Epub

2008 Jan 2.

Prediction of protein functional residues from sequence by probability density

estimation.

Fischer JD(1), Mayer CE, Söding J.

Author information:

(1)Department for Protein Evolution, Max Planck Institute for Developmental

Biology, Spemannstr. 35, 72076 Tübingen, Germany.

MOTIVATION: The prediction of ligand-binding residues or catalytically active

residues of a protein may give important hints that can guide further genetic or

biochemical studies. Existing sequence-based prediction methods mostly rank

residue positions by evolutionary conservation calculated from a multiple

sequence alignment of homologs. A problem hampering more wide-spread application

of these methods is the low per-residue precision, which at 20% sensitivity is

around 35% for ligand-binding residues and 20% for catalytic residues.

RESULTS: We combine information from the conservation at each site, its amino

acid distribution, as well as its predicted secondary structure (ss) and relative

solvent accessibility (rsa). First, we measure conservation by how much the amino

acid distribution at each site differs from the distribution expected for the

predicted ss and rsa states. Second, we include the conservation of neighboring

residues in a weighted linear score by analytically optimizing the

signal-to-noise ratio of the total score. Third, we use conditional probability

density estimation to calculate the probability of each site to be functional

given its conservation, the observed amino acid distribution, and the predicted

ss and rsa states. We have constructed two large data sets, one based on the

Catalytic Site Atlas and the other on PDB SITE records, to benchmark methods for

predicting functional residues. The new method FRcons predicts ligand-binding and

catalytic residues with higher precision than alternative methods over the entire

sensitivity range, reaching 50% and 40% precision at 20% sensitivity,

respectively.

AVAILABILITY: Server: http://frpred.tuebingen.mpg.de. Data sets:

ftp://ftp.tuebingen.mpg.de/pub/protevo/FRpred/.

DOI: 10.1093/bioinformatics/btm626

PMID: 18174181 [Indexed for MEDLINE]

2501. Nucleic Acids Res. 2008 Mar;36(5):e34. doi: 10.1093/nar/gkn083. Epub 2008 Feb 22.

Bisulfite sequencing Data Presentation and Compilation (BDPC) web server--a

useful tool for DNA methylation analysis.

Rohde C(1), Zhang Y, Jurkowski TP, Stamerjohanns H, Reinhardt R, Jeltsch A.

Author information:

(1)School of Engineering and Science, Jacobs University Bremen, Campus Ring 1,

28725 Bremen, Germany.

During bisulfite genomic sequencing projects large amount of data are generated.

The Bisulfite sequencing Data Presentation and Compilation (BDPC) web interface

(http://biochem.jacobs-university.de/BDPC/) automatically analyzes bisulfite

datasets prepared using the BiQ Analyzer. BDPC provides the following output: (i)

MS-Excel compatible files compiling for each PCR product (a) the average

methylation level, the number of clones analyzed and the percentage of CG sites

analyzed (which is an indicator of data quality), (b) the methylation level

observed at each CG site and (c) the methylation level of each clone. (ii) A

methylation overview table compiling the methylation of all amplicons in all

tissues. (iii) Publication grade figures in PNG format showing the methylation

pattern for each PCR product embedded in an HMTL file summarizing the methylation

data, the DNA sequence and some basic statistics. (iv) A summary file compiling

the methylation pattern of different tissues, which is linked to the individual

HTML result files, and can be directly used for presentation of the data in the

internet. (v) A condensed file, containing all primary data in simplified format

for further downstream data analysis and (vi) a custom track file for display of

the results in the UCSC genome browser.

DOI: 10.1093/nar/gkn083

PMCID: PMC2275153

PMID: 18296484 [Indexed for MEDLINE]

2502. Proteins. 2008 Mar;70(4):1219-27.

HingeProt: automated prediction of hinges in protein structures.

Emekli U(1), Schneidman-Duhovny D, Wolfson HJ, Nussinov R, Haliloglu T.

Author information:

(1)Polymer Research Center and Chemical Engineering Department, Bogaziçi

University, 34342 Bebek, Istanbul, Turkey.

Proteins are highly flexible molecules. Prediction of molecular flexibility aids

in the comprehension and prediction of protein function and in providing details

of functional mechanisms. The ability to predict the locations, directions, and

extent of molecular movements can assist in fitting atomic resolution structures

to low-resolution EM density maps and in predicting the complex structures of

interacting molecules (docking). There are several types of molecular movements.

In this work, we focus on the prediction of hinge movements. Given a single

protein structure, the method automatically divides it into the rigid parts and

the hinge regions connecting them. The method employs the Elastic Network Model,

which is very efficient and was validated against a large data set of proteins.

The output can be used in applications such as flexible protein-protein and

protein-ligand docking, flexible docking of protein structures into cryo-EM maps,

and refinement of low-resolution EM structures. The web server of HingeProt

provides convenient visualization of the results and is available with two mirror

sites at http://www.prc.boun.edu.tr/appserv/prc/HingeProt3 and

http://bioinfo3d.cs.tau.ac.il/HingeProt/.

2007 Wiley-Liss, Inc.

DOI: 10.1002/prot.21613

PMID: 17847101 [Indexed for MEDLINE]

2503. Trends Biochem Sci. 2008 Mar;33(3):101-3. doi: 10.1016/j.tibs.2008.01.001. Epub

2008 Feb 13.

iPath: interactive exploration of biochemical pathways and networks.

Letunic I(1), Yamada T, Kanehisa M, Bork P.

Author information:

(1)EMBL, Meyerhofstrasse 1, Heidelberg, Germany.

iPath is an open-access online tool (http://pathways.embl.de) for visualizing and

analyzing metabolic pathways. An interactive viewer provides straightforward

navigation through various pathways and enables easy access to the underlying

chemicals and enzymes. Customized pathway maps can be generated and annotated

using various external data. For example, by merging human genome data with two

important gut commensals, iPath can pinpoint the complementarity of the

host-symbiont metabolic capacities.

DOI: 10.1016/j.tibs.2008.01.001

PMID: 18276143 [Indexed for MEDLINE]

2504. BMC Bioinformatics. 2008 Feb 22;9:112. doi: 10.1186/1471-2105-9-112.

ReAlignerV: web-based genomic alignment tool with high specificity and robustness

estimated by species-specific insertion sequences.

Iwama H(1), Hori Y, Matsumoto K, Murao K, Ishida T.

Author information:

(1)Life Science Research Center, Kagawa University, Ikenobe 1750-1, Miki-cho,

Kita-gun, Kagawa, 761-0793, Japan. iwama@med.kagawa-u.ac.jp

BACKGROUND: Detecting conserved noncoding sequences (CNSs) across species

highlights the functional elements. Alignment procedures combined with

computational prediction of transcription factor binding sites (TFBSs) can narrow

down key regulatory elements. Repeat masking processes are often performed before

alignment to mask insertion sequences such as transposable elements (TEs).

However, recently such TEs have been reported to influence the gene regulatory

network evolution. Therefore, an alignment approach that is robust to TE

insertions is meaningful for finding novel conserved TFBSs in TEs.

RESULTS: We constructed a web server 'ReAlignerV' for complex alignment of

genomic sequences. ReAlignerV returns ladder-like schematic alignments that

integrate predicted TFBSs and the location of TEs. It also provides pair-wise

alignments in which the predicted TFBS sites and their names are shown alongside

each sequence. Furthermore, we evaluated false positive aligned sites by focusing

on the species-specific TEs (SSTEs), and found that ReAlignerV has a higher

specificity and robustness to insertions for sequences having more than 20% TE

content, compared to LAGAN, AVID, MAVID and BLASTZ.

CONCLUSION: ReAlignerV can be applied successfully to TE-insertion-rich sequences

without prior repeat masking, and this increases the chances of finding

regulatory sequences hidden in TEs, which are important sources of the regulatory

network evolution. ReAlignerV can be accessed through and downloaded from

http://genet.med.kagawa-u.ac.jp/.

DOI: 10.1186/1471-2105-9-112

PMCID: PMC2267439

PMID: 18294369 [Indexed for MEDLINE]

2505. PLoS One. 2008 Feb 20;3(2):e1623. doi: 10.1371/journal.pone.0001623.

Asap: a framework for over-representation statistics for transcription factor

binding sites.

Marstrand TT(1), Frellsen J, Moltke I, Thiim M, Valen E, Retelska D, Krogh A.

Author information:

(1)Bioinformatics Centre, Department of Molecular Biology and Biotech Research

and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark.

BACKGROUND: In studies of gene regulation the efficient computational detection

of over-represented transcription factor binding sites is an increasingly

important aspect. Several published methods can be used for testing whether a set

of hypothesised co-regulated genes share a common regulatory regime based on the

occurrence of the modelled transcription factor binding sites. However there is

little or no information available for guiding the end users choice of method.

Furthermore it would be necessary to obtain several different software programs

from various sources to make a well-founded choice.

METHODOLOGY: We introduce a software package, Asap, for fast searching with

position weight matrices that include several standard methods for assessing

over-representation. We have compared the ability of these methods to detect

over-represented transcription factor binding sites in artificial promoter

sequences. Controlling all aspects of our input data we are able to identify the

optimal statistics across multiple threshold values and for sequence sets

containing different distributions of transcription factor binding sites.

CONCLUSIONS: We show that our implementation is significantly faster than more

naïve scanning algorithms when searching with many weight matrices in large

sequence sets. When comparing the various statistics, we show that those based on

binomial over-representation and Fisher's exact test performs almost equally good

and better than the others. An online server is available at

http://servers.binf.ku.dk/asap/.

DOI: 10.1371/journal.pone.0001623

PMCID: PMC2229843

PMID: 18286180 [Indexed for MEDLINE]

2506. BMC Bioinformatics. 2008 Feb 19;9:104. doi: 10.1186/1471-2105-9-104.

PDTD: a web-accessible protein database for drug target identification.

Gao Z(1), Li H, Zhang H, Liu X, Kang L, Luo X, Zhu W, Chen K, Wang X, Jiang H.

Author information:

(1)Drug Discovery and Design Center, State Key Laboratory of Drug Research,

Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai

201203, China. zhentg@mail.shcnc.ac.cn

BACKGROUND: Target identification is important for modern drug discovery. With

the advances in the development of molecular docking, potential binding proteins

may be discovered by docking a small molecule to a repository of proteins with

three-dimensional (3D) structures. To complete this task, a reverse docking

program and a drug target database with 3D structures are necessary. To this end,

we have developed a web server tool, TarFisDock (Target Fishing Docking)

http://www.dddc.ac.cn/tarfisdock, which has been used widely by others. Recently,

we have constructed a protein target database, Potential Drug Target Database

(PDTD), and have integrated PDTD with TarFisDock. This combination aims to assist

target identification and validation.

DESCRIPTION: PDTD is a web-accessible protein database for in silico target

identification. It currently contains >1100 protein entries with 3D structures

presented in the Protein Data Bank. The data are extracted from the literatures

and several online databases such as TTD, DrugBank and Thomson Pharma. The

database covers diverse information of >830 known or potential drug targets,

including protein and active sites structures in both PDB and mol2 formats,

related diseases, biological functions as well as associated regulating

(signaling) pathways. Each target is categorized by both nosology and biochemical

function. PDTD supports keyword search function, such as PDB ID, target name, and

disease name. Data set generated by PDTD can be viewed with the plug-in of

molecular visualization tools and also can be downloaded freely. Remarkably, PDTD

is specially designed for target identification. In conjunction with TarFisDock,

PDTD can be used to identify binding proteins for small molecules. The results

can be downloaded in the form of mol2 file with the binding pose of the probe

compound and a list of potential binding targets according to their ranking

scores.

CONCLUSION: PDTD serves as a comprehensive and unique repository of drug targets.

Integrated with TarFisDock, PDTD is a useful resource to identify binding

proteins for active compounds or existing drugs. Its potential applications

include in silico drug target identification, virtual screening, and the

discovery of the secondary effects of an old drug (i.e. new pharmacological

usage) or an existing target (i.e. new pharmacological or toxic relevance), thus

it may be a valuable platform for the pharmaceutical researchers. PDTD is

available online at http://www.dddc.ac.cn/pdtd/.

DOI: 10.1186/1471-2105-9-104

PMCID: PMC2265675

PMID: 18282303 [Indexed for MEDLINE]

2507. Anal Biochem. 2008 Feb 15;373(2):386-8. Epub 2007 Oct 13.

PseAAC: a flexible web server for generating various kinds of protein pseudo

amino acid composition.

Shen HB(1), Chou KC.

Author information:

(1)Gordon Life Science Institute, San Diego, CA 92130, USA.

hbshen@crystal.harvard.edu

The pseudo amino acid (PseAA) composition can represent a protein sequence in a

discrete model without completely losing its sequence-order information, and

hence has been widely applied for improving the prediction quality for various

protein attributes. However, dealing with different problems may need different

kinds of PseAA composition. Here, we present a web-server called PseAAC at

http://chou.med.harvard.edu/bioinf/PseAA/, by which users can generate various

kinds of PseAA composition to best fit their need.

DOI: 10.1016/j.ab.2007.10.012

PMID: 17976365 [Indexed for MEDLINE]

2508. Bioinformatics. 2008 Feb 15;24(4):586-7. doi: 10.1093/bioinformatics/btn014. Epub

2008 Jan 9.

The ModFOLD server for the quality assessment of protein structural models.

McGuffin LJ(1).

Author information:

(1)School of Biological Sciences, University of Reading, Whiteknights, Reading

RG6 6AS, UK. l.j.mcguffin@reading.ac.uk

The reliable assessment of the quality of protein structural models is

fundamental to the progress of structural bioinformatics. The ModFOLD server

provides access to two accurate techniques for the global and local prediction of

the quality of 3D models of proteins. Firstly ModFOLD, which is a fast Model

Quality Assessment Program (MQAP) used for the global assessment of either single

or multiple models. Secondly ModFOLDclust, which is a more intensive method that

carries out clustering of multiple models and provides per-residue local quality

assessment.AVAILABILITY: http://www.biocentre.rdg.ac.uk/bioinformatics/ModFOLD/.

DOI: 10.1093/bioinformatics/btn014

PMID: 18184684 [Indexed for MEDLINE]

2509. Bioinformatics. 2008 Feb 15;24(4):577-8. doi: 10.1093/bioinformatics/btm594. Epub

2008 Jan 6.

FAST: Fourier transform based algorithms for significance testing of ungapped

multiple alignments.

Nagarajan N(1), Keich U.

Author information:

(1)University of Maryland, College Park, MD 20740, USA.

SUMMARY: As was shown in Nagarajan et al. (2005), commonly used approximations

for assessing the significance of multiple alignments can be be very inaccurate.

To address this, we present here the FAST package, an open-source collection of

programs and libraries for efficiently and reliably computing the significance of

ungapped local alignments. We also describe other potential applications in

Bioinformatics where these programs can be adapted for significance testing.

AVAILABILITY: The FAST package includes C++ implementations of various algorithms

that can be used as stand-alone programs or as a library of subroutines. The

package and a web-server for some of the programs are available at

www.cs.cornell.edu/~keich/FAST.

DOI: 10.1093/bioinformatics/btm594

PMID: 18180239 [Indexed for MEDLINE]

2510. Bioinformatics. 2008 Feb 15;24(4):594-6. doi: 10.1093/bioinformatics/btm630. Epub

2008 Jan 2.

NetworkBLAST: comparative analysis of protein networks.

Kalaev M(1), Smoot M, Ideker T, Sharan R.

Author information:

(1)School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel.

kalaevma@post.tau.ac.il

The identification of protein complexes is a fundamental challenge in

interpreting protein-protein interaction data. Cross-species analysis allows

coping with the high levels of noise that are typical to these data. The

NetworkBLAST web-server provides a platform for identifying protein complexes in

protein-protein interaction networks. It can analyze a single network or two

networks from different species. In the latter case, NetworkBLAST outputs a set

of putative complexes that are evolutionarily conserved across the two

networks.AVAILABILITY: NetworkBLAST is available as web-server at:

www.cs.tau.ac.il/~roded/networkblast.htm.

DOI: 10.1093/bioinformatics/btm630

PMID: 18174180 [Indexed for MEDLINE]

2511. Proteins. 2008 Feb 15;70(3):659-66.

MAAP: malarial adhesins and adhesin-like proteins predictor.

Ansari FA(1), Kumar N, Bala Subramanyam M, Gnanamani M, Ramachandran S.

Author information:

(1)G. N. Ramachandran Knowledge Centre for Genome Informatics, Institute of

Genomics and Integrative Biology, Delhi 110 007, India.

Malaria caused by protozoan parasites belonging to the genus Plasmodium is a

dreaded disease, second only to tuberculosis. The emergence of parasites

resistant to commonly used drugs and the lack of availability of vaccines

aggravates the problem. One of the preventive approaches targets adhesion of

parasites to host cells and tissues. Adhesion of parasites is mediated by

proteins called adhesins. Abrogation of adhesion by either immunizing the host

with adhesins or inhibiting the interaction using structural analogs of host cell

receptors holds the potential to develop novel preventive strategies. The

availability of complete genome sequence offers new opportunities for identifying

adhesin and adhesin-like proteins. Development of computational algorithms can

simplify this task and accelerate experimental characterization of the predicted

adhesins from complete genomes. A curated positive dataset of experimentally

known adhesins from Plasmodium species was prepared by careful examination of

literature reports. "Controversial" or "hypothetical" adhesins were excluded. The

negative dataset consisted of proteins representing various intracellular

functions including information processing, metabolism, and interface

(transporters). We did not include proteins likely to be on the surface with

unknown adhesin properties or which are linked even indirectly to the adhesion

process in either of the training sets. A nonhomology-based approach using 420

compositional properties of amino acid dipeptide and multiplet frequencies was

used to develop MAAP Web server with Support Vector Machine (SVM) model

classifier as its engine for the prediction of malarial adhesins and adhesin-like

proteins. The MAAP engine has six SVM classifier models identified through an

exhaustive search from 728 kernel parameters set. These models displayed an

efficiency (Mathews correlation coefficient) of 0.860-0.967. The final prediction

P(maap) score is the maximum score attained by a given sequence in any of the six

models. The results of MAAP runs on complete proteomes of Plasmodium species

revealed that in Plasmodium falciparum at P(maap) scores above 0.0, we observed a

sensitivity of 100% with two false positives. In P. vivax and P. yoelii an

optimal threshold P(maap) score of 0.7 was optimal with very few false positives

(upto 5). Several new predictions were obtained. This list includes hypothetical

protein PF14\_0040, interspersed repeat antigen, STEVOR, liver stage antigen,

SURFIN, RIFIN, stevor (3D7-stevorT3-2), mature parasite-infected erythrocyte

surface antigen or P. falciparum erythrocyte membrane protein 2, merozoite

surface protein 6 in P. falciparum, circumsporozoite proteins, microneme

protein-1, Vir18, Vir12-like, Vir12, Vir18-like, Vir18-related and Vir4 in P.

vivax, circumsporozoite protein/thrombospondin related anonymous proteins, 28 kDa

ookinete surface protein, yir1, and yir4 of P. yoelii. Among these, a few

proteins identified by MAAP were matched with those identified by other groups

using different experimental and theoretical strategies. Most other interspersed

repeat proteins in Plasmodium species had lower P(maap) scores. These new

predictions could serve as new leads for further experimental characterization

(MAAP webserver: http://maap.igib.res.in).

(c) 2007 Wiley-Liss, Inc.

DOI: 10.1002/prot.21568

PMID: 17879344 [Indexed for MEDLINE]

2512. Bioinform Biol Insights. 2008 Feb 1;2:5-13.

Structural re-alignment in an immunogenic surface region of ricin A chain.

Zemla AT(1), Ecale Zhou CL.

Author information:

(1)Computational Biology for Countermeasures Group, Lawrence Livermore National

Laboratory, Livermore, CA 94550, USA.

We compared structure alignments generated by several protein structure

comparison programs to determine whether existing methods would satisfactorily

align residues at a highly conserved position within an immunogenic loop in

ribosome inactivating proteins (RIPs). Using default settings, structure

alignments generated by several programs (CE, DaliLite, FATCAT, LGA, MAMMOTH,

MATRAS, SHEBA, SSM) failed to align the respective conserved residues, although

LGA reported correct residue-residue (R-R) correspondences when the beta-carbon

(Cb) position was used as the point of reference in the alignment calculations.

Further tests using variable points of reference indicated that points distal

from the beta carbon along a vector connecting the alpha and beta carbons yielded

rigid structural alignments in which residues known to be highly conserved in

RIPs were reported as corresponding residues in structural comparisons between

ricin A chain, abrin-A, and other RIPs. Results suggest that approaches to

structure alignment employing alternate point representations corresponding to

side chain position may yield structure alignments that are more consistent with

observed conservation of functional surface residues than do standard alignment

programs, which apply uniform criteria for alignment (i.e. alpha carbon (Ca) as

point of reference) along the entirety of the peptide chain. We present the

results of tests that suggest the utility of allowing user-specified points of

reference in generating alternate structural alignments, and we present a web

server for automatically generating such alignments:

http://as2ts.llnl.gov/AS2TS/LGA/lga\_pdblist\_plots.html.

PMCID: PMC2735970

PMID: 19812763

2513. Bioinformatics. 2008 Feb 1;24(3):358-66. Epub 2007 Dec 14.

Efficient peptide-MHC-I binding prediction for alleles with few known binders.

Jacob L(1), Vert JP.

Author information:

(1)Centre for Computational Biology, Ecole des Mines de Paris, 35 rue Saint

Honoré, 77305 Fontainebleau Cedex, France. laurent.jacob@ensmp.fr

MOTIVATION: In silico methods for the prediction of antigenic peptides binding to

MHC class I molecules play an increasingly important role in the identification

of T-cell epitopes. Statistical and machine learning methods in particular are

widely used to score candidate binders based on their similarity with known

binders and non-binders. The genes coding for the MHC molecules, however, are

highly polymorphic, and statistical methods have difficulties building models for

alleles with few known binders. In this context, recent work has demonstrated the

utility of leveraging information across alleles to improve the performance of

the prediction.

RESULTS: We design a support vector machine algorithm that is able to learn

peptide-MHC-I binding models for many alleles simultaneously, by sharing binding

information across alleles. The sharing of information is controlled by a

user-defined measure of similarity between alleles. We show that this similarity

can be defined in terms of supertypes, or more directly by comparing key residues

known to play a role in the peptide-MHC binding. We illustrate the potential of

this approach on various benchmark experiments where it outperforms other

state-of-the-art methods.

AVAILABILITY: The method is implemented on a web server:

http://cbio.ensmp.fr/kiss. All data and codes are freely and publicly available

from the authors.

DOI: 10.1093/bioinformatics/btm611

PMID: 18083718 [Indexed for MEDLINE]

2514. Bioinformatics. 2008 Feb 1;24(3):424-5. Epub 2007 Dec 1.

BAGET: a web server for the effortless retrieval of prokaryotic gene context and

sequence.

Oberto J(1).

Author information:

(1)Biologie Moléculaire du Gène chez les Extrêmophiles, Université Paris-Sud,

Institut de Génétique et Microbiologie, CNRS UMR8621, 91405 Orsay Cedex, France.

jacques.oberto@igmors.u-psud.fr

BAGET (Bacterial and Archaeal Gene Exploration Tool) is a web service designed to

facilitate extraction, by molecular geneticists and phylogeneticists, of specific

gene and protein sequences from completely determined prokaryotic genomes. Upon

selection of a particular prokaryotic organism and gene, two levels of visual

gene context information are provided on a single dynamic page: (i) a graphical

representation of a user defined portion of the chromosome centered on the gene

of interest and (ii) the DNA sequence of the query gene, of the immediate

neighboring genes and the intergenic regions each identified by a consistent

color code. The aminoacid sequence is provided for protein-coding query genes.

Query results can be exported as a rich text format (RTF) word processor file for

printing, archival or further analysis.AVAILABILITY:

http://archaea.u-psud.fr/bin/baget.dll.

DOI: 10.1093/bioinformatics/btm600

PMID: 18056064 [Indexed for MEDLINE]

2515. J Mol Graph Model. 2008 Feb;26(6):957-61. Epub 2007 Jul 28.

Toward the atomistic simulation of T cell epitopes automated construction of MHC:

peptide structures for free energy calculations.

Todman SJ(1), Halling-Brown MD, Davies MN, Flower DR, Kayikci M, Moss DS.

Author information:

(1)Department of Crystallography, University of London, Birkbeck College, Malet

Street, London WC1E 7HX, United Kingdom.

Epitopes mediated by T cells lie at the heart of the adaptive immune response and

form the essential nucleus of anti-tumour peptide or epitope-based vaccines.

Antigenic T cell epitopes are mediated by major histocompatibility complex (MHC)

molecules, which present them to T cell receptors. Calculating the affinity

between a given MHC molecule and an antigenic peptide using experimental

approaches is both difficult and time consuming, thus various computational

methods have been developed for this purpose. A server has been developed to

allow a structural approach to the problem by generating specific MHC:peptide

complex structures and providing configuration files to run molecular modelling

simulations upon them. A system has been produced which allows the automated

construction of MHC:peptide structure files and the corresponding configuration

files required to execute a molecular dynamics simulation using NAMD. The system

has been made available through a web-based front end and stand-alone scripts.

Previous attempts at structural prediction of MHC:peptide affinity have been

limited due to the paucity of structures and the computational expense in running

large scale molecular dynamics simulations. The MHCsim server

(http://igrid-ext.cryst.bbk.ac.uk/MHCsim) allows the user to rapidly generate any

desired MHC:peptide complex and will facilitate molecular modelling simulation of

MHC complexes on an unprecedented scale.

DOI: 10.1016/j.jmgm.2007.07.005

PMID: 17766153 [Indexed for MEDLINE]

2516. Mol Divers. 2008 Feb;12(1):41-5. doi: 10.1007/s11030-008-9073-0. Epub 2008 May

28.

Using AdaBoost for the prediction of subcellular location of prokaryotic and

eukaryotic proteins.

Niu B(1), Jin YH, Feng KY, Lu WC, Cai YD, Li GZ.

Author information:

(1)School of Materials Science and Engineering, Shanghai University, Shanghai,

200444, China.

In this paper, AdaBoost algorithm, a popular and effective prediction method, is

applied to predict the subcellular locations of Prokaryotic and Eukaryotic

Proteins-a dataset derived from SWISSPROT 33.0. Its prediction ability was

evaluated by re-substitution test, Leave-One-Out Cross validation (LOOCV) and

jackknife test. By comparing its results with some most popular predictors such

as Discriminant Function, neural networks, and SVM, we demonstrated that the

AdaBoost predictor outperformed these predictors. As a result, we arrive at the

conclusion that AdaBoost algorithm could be employed as a robust method to

predict subcellular location. An online web server for predicting subcellular

location of prokaryotic and eukaryotic proteins is available at

http://chemdata.shu.edu.cn/subcell/ .

DOI: 10.1007/s11030-008-9073-0

PMID: 18506593 [Indexed for MEDLINE]

2517. BMC Bioinformatics. 2008 Jan 29;9:65. doi: 10.1186/1471-2105-9-65.

E-CAI: a novel server to estimate an expected value of Codon Adaptation Index

(eCAI).

Puigbò P(1), Bravo IG, Garcia-Vallvé S.

Author information:

(1)Evolutionary Genomics Group, Department of Biochemistry and Biotechnology,

Rovira i Virgili University (URV), Campus Sescelades, c/Marcelli Domingo s/n,

43007 Tarragona, Spain. ppuigbo@urv.cat

BACKGROUND: The Codon Adaptation Index (CAI) is a measure of the synonymous codon

usage bias for a DNA or RNA sequence. It quantifies the similarity between the

synonymous codon usage of a gene and the synonymous codon frequency of a

reference set. Extreme values in the nucleotide or in the amino acid composition

have a large impact on differential preference for synonymous codons. It is

thence essential to define the limits for the expected value of CAI on the basis

of sequence composition in order to properly interpret the CAI and provide

statistical support to CAI analyses. Though several freely available programs

calculate the CAI for a given DNA sequence, none of them corrects for

compositional biases or provides confidence intervals for CAI values.

RESULTS: The E-CAI server, available at http://genomes.urv.es/CAIcal/E-CAI, is a

web-application that calculates an expected value of CAI for a set of query

sequences by generating random sequences with G+C and amino acid content similar

to those of the input. An executable file, a tutorial, a Frequently Asked

Questions (FAQ) section and several examples are also available. To exemplify the

use of the E-CAI server, we have analysed the codon adaptation of human

mitochondrial genes that codify a subunit of the mitochondrial respiratory chain

(excluding those genes that lack a prokaryotic orthologue) and are encoded in the

nuclear genome. It is assumed that these genes were transferred from the

proto-mitochondrial to the nuclear genome and that its codon usage was then

ameliorated.

CONCLUSION: The E-CAI server provides a direct threshold value for discerning

whether the differences in CAI are statistically significant or whether they are

merely artifacts that arise from internal biases in the G+C composition and/or

amino acid composition of the query sequences.

DOI: 10.1186/1471-2105-9-65

PMCID: PMC2246156

PMID: 18230160 [Indexed for MEDLINE]

2518. BMC Bioinformatics. 2008 Jan 28;9:62. doi: 10.1186/1471-2105-9-62.

VirulentPred: a SVM based prediction method for virulent proteins in bacterial

pathogens.

Garg A(1), Gupta D.

Author information:

(1)Structural and Computational Biology Group, International Centre for Genetic

Engineering and Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhi 110067,

India. aarti@icgeb.res.in

BACKGROUND: Prediction of bacterial virulent protein sequences has implications

for identification and characterization of novel virulence-associated factors,

finding novel drug/vaccine targets against proteins indispensable to

pathogenicity, and understanding the complex virulence mechanism in pathogens.

RESULTS: In the present study we propose a bacterial virulent protein prediction

method based on bi-layer cascade Support Vector Machine (SVM). The first layer

SVM classifiers were trained and optimized with different individual protein

sequence features like amino acid composition, dipeptide composition (occurrences

of the possible pairs of ith and i+1th amino acid residues), higher order

dipeptide composition (pairs of ith and i+2nd residues) and Position Specific

Iterated BLAST (PSI-BLAST) generated Position Specific Scoring Matrices (PSSM).

In addition, a similarity-search based module was also developed using a dataset

of virulent and non-virulent proteins as BLAST database. A five-fold

cross-validation technique was used for the evaluation of various prediction

strategies in this study. The results from the first layer (SVM scores and

PSI-BLAST result) were cascaded to the second layer SVM classifier to train and

generate the final classifier. The cascade SVM classifier was able to accomplish

an accuracy of 81.8%, covering 86% area in the Receiver Operator Characteristic

(ROC) plot, better than that of either of the layer one SVM classifiers based on

single or multiple sequence features.

CONCLUSION: VirulentPred is a SVM based method to predict bacterial virulent

proteins sequences, which can be used to screen virulent proteins in proteomes.

Together with experimentally verified virulent proteins, several putative, non

annotated and hypothetical protein sequences have been predicted to be high

scoring virulent proteins by the prediction method. VirulentPred is available as

a freely accessible World Wide Web server - VirulentPred, at

http://bioinfo.icgeb.res.in/virulent/.

DOI: 10.1186/1471-2105-9-62

PMCID: PMC2254373

PMID: 18226234 [Indexed for MEDLINE]

2519. BMC Bioinformatics. 2008 Jan 25;9:53. doi: 10.1186/1471-2105-9-53.

TOBFAC: the database of tobacco transcription factors.

Rushton PJ(1), Bokowiec MT, Laudeman TW, Brannock JF, Chen X, Timko MP.

Author information:

(1)Department of Biology, University of Virginia, Charlottesville, VA 22904, USA.

pr8y@virginia.edu

BACKGROUND: Regulation of gene expression at the level of transcription is a

major control point in many biological processes. Transcription factors (TFs) can

activate and/or repress the transcriptional rate of target genes and vascular

plant genomes devote approximately 7% of their coding capacity to TFs. Global

analysis of TFs has only been performed for three complete higher plant genomes -

Arabidopsis (Arabidopsis thaliana), poplar (Populus trichocarpa) and rice (Oryza

sativa). Presently, no large-scale analysis of TFs has been made from a member of

the Solanaceae, one of the most important families of vascular plants. To fill

this void, we have analysed tobacco (Nicotiana tabacum) TFs using a dataset of

1,159,022 gene-space sequence reads (GSRs) obtained by methylation filtering of

the tobacco genome. An analytical pipeline was developed to isolate TF sequences

from the GSR data set. This involved multiple (typically 10-15) independent

searches with different versions of the TF family-defining domain(s) (normally

the DNA-binding domain) followed by assembly into contigs and verification. Our

analysis revealed that tobacco contains a minimum of 2,513 TFs representing all

of the 64 well-characterised plant TF families. The number of TFs in tobacco is

higher than previously reported for Arabidopsis and rice.

RESULTS: TOBFAC: the database of tobacco transcription factors, is an integrative

database that provides a portal to sequence and phylogeny data for the identified

TFs, together with a large quantity of other data concerning TFs in tobacco. The

database contains an individual page dedicated to each of the 64 TF families.

These contain background information, domain architecture via Pfam links, a list

of all sequences and an assessment of the minimum number of TFs in this family in

tobacco. Downloadable phylogenetic trees of the major families are provided along

with detailed information on the bioinformatic pipeline that was used to find all

family members. TOBFAC also contains EST data, a list of published tobacco TFs

and a list of papers concerning tobacco TFs. The sequences and annotation data

are stored in relational tables using a PostgrelSQL relational database

management system. The data processing and analysis pipelines used the Perl

programming language. The web interface was implemented in JavaScript and Perl

CGI running on an Apache web server. The computationally intensive data

processing and analysis pipelines were run on an Apple XServe cluster with more

than 20 nodes.

CONCLUSION: TOBFAC is an expandable knowledgebase of tobacco TFs with data

currently available for over 2,513 TFs from 64 gene families. TOBFAC integrates

available sequence information, phylogenetic analysis, and EST data with

published reports on tobacco TF function. The database provides a major resource

for the study of gene expression in tobacco and the Solanaceae and helps to fill

a current gap in studies of TF families across the plant kingdom. TOBFAC is

publicly accessible at http://compsysbio.achs.virginia.edu/tobfac/.

DOI: 10.1186/1471-2105-9-53

PMCID: PMC2246155

PMID: 18221524 [Indexed for MEDLINE]

2520. BMC Bioinformatics. 2008 Jan 24;9:41. doi: 10.1186/1471-2105-9-41.

Predicting the interactome of Xanthomonas oryzae pathovar oryzae for target

selection and DB service.

Kim JG(1), Park D, Kim BC, Cho SW, Kim YT, Park YJ, Cho HJ, Park H, Kim KB, Yoon

KO, Park SJ, Lee BM, Bhak J.

Author information:

(1)Microbial Genetics Division, National Institute of Agricultural Biotechnology

(NIAB), Rural Development Administration (RDA), Suwon 441-707, Korea.

jkim5aug@yahoo.com

BACKGROUND: Protein-protein interactions (PPIs) play key roles in various

cellular functions. In addition, some critical inter-species interactions such as

host-pathogen interactions and pathogenicity occur through PPIs. Phytopathogenic

bacteria infect hosts through attachment to host tissue, enzyme secretion,

exopolysaccharides production, toxins release, iron acquisition, and effector

proteins secretion. Many such mechanisms involve some kind of protein-protein

interaction in hosts. Our first aim was to predict the whole protein interaction

pairs (interactome) of Xanthomonas oryzae pathovar oryzae (Xoo) that is an

important pathogenic bacterium that causes bacterial blight (BB) in rice. We

developed a detection protocol to find possibly interacting proteins in its host

using whole genome PPI prediction algorithms. The second aim was to build a DB

server and a bioinformatic procedure for finding target proteins in Xoo for

developing pesticides that block host-pathogen protein interactions within

critical biochemical pathways.

DESCRIPTION: A PPI network in Xoo proteome was predicted by bioinformatics

algorithms: PSIMAP, PEIMAP, and iPfam. We present the resultant species specific

interaction network and host-pathogen interaction, XooNET. It is a comprehensive

predicted initial PPI data for Xoo. XooNET can be used by experimentalists to

pick up protein targets for blocking pathological interactions. XooNET uses most

of the major types of PPI algorithms. They are: 1) Protein Structural Interactome

MAP (PSIMAP), a method using structural domain of SCOP, 2) Protein Experimental

Interactome MAP (PEIMAP), a common method using public resources of experimental

protein interaction information such as HPRD, BIND, DIP, MINT, IntAct, and

BioGrid, and 3) Domain-domain interactions, a method using Pfam domains such as

iPfam. Additionally, XooNET provides information on network properties of the Xoo

interactome.

CONCLUSION: XooNET is an open and free public database server for protein

interaction information for Xoo. It contains 4,538 proteins and 26,932 possible

interactions consisting of 18,503 (PSIMAP), 3,118 (PEIMAP), and 8,938 (iPfam)

pairs. In addition, XooNET provides 3,407 possible interaction pairs between two

sets of proteins; 141 Xoo proteins that are predicted as membrane proteins and

rice proteomes. The resultant interacting partners of a query protein can be

easily retrieved by users as well as the interaction networks in graphical web

interfaces. XooNET is freely available from http://bioportal.kobic.kr/XooNET/.

DOI: 10.1186/1471-2105-9-41

PMCID: PMC2246157

PMID: 18215330 [Indexed for MEDLINE]

2521. BMC Bioinformatics. 2008 Jan 23;9:40. doi: 10.1186/1471-2105-9-40.

I-TASSER server for protein 3D structure prediction.

Zhang Y(1).

Author information:

(1)Center for Bioinformatics and Department of Molecular Bioscience, University

of Kansas, 2030 Becker Dr, Lawrence, KS 66047, USA. yzhang@ku.edu

BACKGROUND: Prediction of 3-dimensional protein structures from amino acid

sequences represents one of the most important problems in computational

structural biology. The community-wide Critical Assessment of Structure

Prediction (CASP) experiments have been designed to obtain an objective

assessment of the state-of-the-art of the field, where I-TASSER was ranked as the

best method in the server section of the recent 7th CASP experiment. Our

laboratory has since then received numerous requests about the public

availability of the I-TASSER algorithm and the usage of the I-TASSER predictions.

RESULTS: An on-line version of I-TASSER is developed at the KU Center for

Bioinformatics which has generated protein structure predictions for thousands of

modeling requests from more than 35 countries. A scoring function (C-score) based

on the relative clustering structural density and the consensus significance

score of multiple threading templates is introduced to estimate the accuracy of

the I-TASSER predictions. A large-scale benchmark test demonstrates a strong

correlation between the C-score and the TM-score (a structural similarity

measurement with values in [0, 1]) of the first models with a correlation

coefficient of 0.91. Using a C-score cutoff > -1.5 for the models of correct

topology, both false positive and false negative rates are below 0.1. Combining

C-score and protein length, the accuracy of the I-TASSER models can be predicted

with an average error of 0.08 for TM-score and 2 A for RMSD.

CONCLUSION: The I-TASSER server has been developed to generate automated

full-length 3D protein structural predictions where the benchmarked scoring

system helps users to obtain quantitative assessments of the I-TASSER models. The

output of the I-TASSER server for each query includes up to five full-length

models, the confidence score, the estimated TM-score and RMSD, and the standard

deviation of the estimations. The I-TASSER server is freely available to the

academic community at http://zhang.bioinformatics.ku.edu/I-TASSER.

DOI: 10.1186/1471-2105-9-40

PMCID: PMC2245901

PMID: 18215316 [Indexed for MEDLINE]

2522. BMC Bioinformatics. 2008 Jan 23;9:39. doi: 10.1186/1471-2105-9-39.

miRNAminer: a tool for homologous microRNA gene search.

Artzi S(1), Kiezun A, Shomron N.

Author information:

(1)Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA

02139, USA. artzi@mit.edu

BACKGROUND: MicroRNAs (miRNAs), present in most metazoans, are small non-coding

RNAs that control gene expression by negatively regulating translation through

binding to the 3'UTR of mRNA transcripts. Previously, experimental and

computational methods were used to construct miRNA gene repositories agreeing

with careful submission guidelines.

RESULTS: An algorithm we developed - miRNAminer - is used for homologous

conserved miRNA gene search in several animal species. Given a search query,

candidate homologs from different species are tested for their known miRNA

properties, such as secondary structure, energy and alignment and conservation,

in order to asses their fidelity. When applying miRNAminer on seven mammalian

species we identified several hundreds of high-confidence homologous miRNAs

increasing the total collection of (miRbase) miRNAs, in these species, by more

than 50%. miRNAminer uses stringent criteria and exhibits high sensitivity and

specificity.

CONCLUSION: We present - miRNAminer - the first web-server for homologous miRNA

gene search in animals. miRNAminer can be used to identify conserved homolog

miRNA genes and can also be used prior to depositing miRNAs in public databases.

miRNAminer is available at http://pag.csail.mit.edu/mirnaminer.

DOI: 10.1186/1471-2105-9-39

PMCID: PMC2258288

PMID: 18215311 [Indexed for MEDLINE]

2523. BMC Bioinformatics. 2008 Jan 23;9:33. doi: 10.1186/1471-2105-9-33.

A fast structural multiple alignment method for long RNA sequences.

Tabei Y(1), Kiryu H, Kin T, Asai K.

Author information:

(1)Graduate School of Frontier Science, University of Tokyo, CB04 Kiban-tou 5-1-5

Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan. tabei@cb.k.u-tokyo.ac.jp

BACKGROUND: Aligning multiple RNA sequences is essential for analyzing non-coding

RNAs. Although many alignment methods for non-coding RNAs, including Sankoff's

algorithm for strict structural alignments, have been proposed, they are either

inaccurate or computationally too expensive. Faster methods with reasonable

accuracies are required for genome-scale analyses.

RESULTS: We propose a fast algorithm for multiple structural alignments of RNA

sequences that is an extension of our pairwise structural alignment method

(implemented in SCARNA). The accuracies of the implemented software, MXSCARNA,

are at least as favorable as those of state-of-art algorithms that are

computationally much more expensive in time and memory.

CONCLUSION: The proposed method for structural alignment of multiple RNA

sequences is fast enough for large-scale analyses with accuracies at least

comparable to those of existing algorithms. The source code of MXSCARNA and its

web server are available at http://mxscarna.ncrna.org.

DOI: 10.1186/1471-2105-9-33

PMCID: PMC2375124

PMID: 18215258 [Indexed for MEDLINE]

2524. BMC Bioinformatics. 2008 Jan 18;9:27. doi: 10.1186/1471-2105-9-27.

T2prhd: a tool to study the patterns of repeat evolution.

Sipos B(1), Somogyi K, Andó I, Pénzes Z.

Author information:

(1)Institute of Genetics, Biological Research Center of the Hungarian Academy of

Sciences, P.O. Box 521, H-6701 Szeged, Hungary. sbotond@gmail.com

BACKGROUND: The models developed to characterize the evolution of multigene

families (such as the birth-and-death and the concerted models) have also been

applied on the level of sequence repeats inside a gene/protein. Phylogenetic

reconstruction is the method of choice to study the evolution of gene families

and also sequence repeats in the light of these models. The characterization of

the gene family evolution in view of the evolutionary models is done by the

evaluation of the clustering of the sequences with the originating loci in mind.

As the locus represents positional information, it is straightforward that in the

case of the repeats the exact position in the sequence should be used, as the

simple numbering according to repeat order can be misleading.

RESULTS: We have developed a novel rapid visual approach to study repeat

evolution, that takes into account the exact repeat position in a sequence. The

"pairwise repeat homology diagram" visualizes sequence repeats detected by a

profile HMM in a pair of sequences and highlights their homology relations

inferred by a phylogenetic tree. The method is implemented in a Perl script

(t2prhd) available for downloading at http://t2prhd.sourceforge.net and is also

accessible as an online tool at http://t2prhd.brc.hu. The power of the method is

demonstrated on the EGF-like and fibronectin-III-like (Fn-III) domain repeats of

three selected mammalian Tenascin sequences.

CONCLUSION: Although pairwise repeat homology diagrams do not carry all the

information provided by the phylogenetic tree, they allow a rapid and intuitive

assessment of repeat evolution. We believe, that t2prhd is a helpful tool with

which to study the pattern of repeat evolution. This method can be particularly

useful in cases of large datasets (such as large gene families), as the command

line interface makes it possible to automate the generation of pairwise repeat

homology diagrams with the aid of scripts.

DOI: 10.1186/1471-2105-9-27

PMCID: PMC2267704

PMID: 18205906 [Indexed for MEDLINE]

2525. Bioinformatics. 2008 Jan 15;24(2):296-8. Epub 2007 Nov 15.

Text processing through Web services: calling Whatizit.

Rebholz-Schuhmann D(1), Arregui M, Gaudan S, Kirsch H, Jimeno A.

Author information:

(1)European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton,

Cambridge, CB10 1SD, UK. rebholz@ebi.ac.uk

MOTIVATION: Text-mining (TM) solutions are developing into efficient services to

researchers in the biomedical research community. Such solutions have to scale

with the growing number and size of resources (e.g. available controlled

vocabularies), with the amount of literature to be processed (e.g. about 17

million documents in PubMed) and with the demands of the user community (e.g.

different methods for fact extraction). These demands motivated the development

of a server-based solution for literature analysis. Whatizit is a suite of

modules that analyse text for contained information, e.g. any scientific

publication or Medline abstracts. Special modules identify terms and then link

them to the corresponding entries in bioinformatics databases such as

UniProtKb/Swiss-Prot data entries and gene ontology concepts. Other modules

identify a set of selected annotation types like the set produced by the EBIMed

analysis pipeline for proteins. In the case of Medline abstracts, Whatizit offers

access to EBI's in-house installation via PMID or term query. For large

quantities of the user's own text, the server can be operated in a streaming mode

(http://www.ebi.ac.uk/webservices/whatizit).

DOI: 10.1093/bioinformatics/btm557

PMID: 18006544 [Indexed for MEDLINE]

2526. Bioinformatics. 2008 Jan 15;24(2):272-5. Epub 2007 Nov 13.

Fast protein fold estimation from NMR-derived distance restraints.

Angyán AF(1), Perczel A, Pongor S, Gáspári Z.

Author information:

(1)Institute of Chemistry, Eötvös Loránd University, Budapest, Hungary.

PRIDE-NMR is a fast novel method to relate known protein folds to NMR distance

restraints. It can be used to obtain a first guess about a structure being

determined, as well as to estimate the completeness or verify the correctness of

NOE data.AVAILABILITY: The PRIDE-NMR server is available at

http://www.icgeb.org/pride

DOI: 10.1093/bioinformatics/btm564

PMID: 18003647 [Indexed for MEDLINE]

2527. Bioinformatics. 2008 Jan 1;24(1):18-25. Epub 2007 Nov 17.

Multi-RELIEF: a method to recognize specificity determining residues from

multiple sequence alignments using a Machine-Learning approach for feature

weighting.

Ye K(1), Feenstra KA, Heringa J, Ijzerman AP, Marchiori E.

Author information:

(1)Division of Medical Chemistry, LACDR, Leiden University, P.O. Box 9502, 2300

RA, Leiden, The Netherlands.

MOTIVATION: Identification of residues that account for protein function

specificity is crucial, not only for understanding the nature of functional

specificity, but also for protein engineering experiments aimed at switching the

specificity of an enzyme, regulator or transporter. Available algorithms

generally use multiple sequence alignments to identify residue positions

conserved within subfamilies but divergent in between. However, many biological

examples show a much subtler picture than simple intra-group conservation versus

inter-group divergence.

RESULTS: We present multi-RELIEF, a novel approach for identifying specificity

residues that is based on RELIEF, a state-of-the-art Machine-Learning technique

for feature weighting. It estimates the expected 'local' functional specificity

of residues from an alignment divided in multiple classes. Optionally, 3D

structure information is exploited by increasing the weight of residues that have

high-weight neighbors. Using ROC curves over a large body of experimental

reference data, we show that (a) multi-RELIEF identifies specificity residues for

the seven test sets used, (b) incorporating structural information improves

prediction for specificity of interaction with small molecules and (c) comparison

of multi-RELIEF with four other state-of-the-art algorithms indicates its

robustness and best overall performance.

AVAILABILITY: A web-server implementation of multi-RELIEF is available at

www.ibi.vu.nl/programs/multirelief. Matlab source code of the algorithm and data

sets are available on request for academic use.

DOI: 10.1093/bioinformatics/btm537

PMID: 18024975 [Indexed for MEDLINE]

2528. Bioinformation. 2008;3(3):137-8. Epub 2008 Nov 9.

ESBRI: a web server for evaluating salt bridges in proteins.

Costantini S(1), Colonna G, Facchiano AM.

Author information:

(1)Laboratory of Bioinformatics and Computational Biology, Institute of Food

Science, CNR, via Roma 52 A/C, 83100 Avellino, Italy.

Salt bridges can play important roles in protein structure and function and have

stabilizing and destabilizing effects in protein folding. ESBRI is a software

available as web tool which analyses the salt bridges in a protein structure,

starting from the atomic coordinates. In the case of protein complexes, the salt

bridges between protein chains can be evaluated, as well as those among specific

charged amino acids and the different protein subunits, in order to obtain useful

information regard the protein-protein interaction.AVAILABILITY: The service is

available at the URL: http://bioinformatica.isa.cnr.it/ESBRI/

PMCID: PMC2639689

PMID: 19238252

2529. Bioinformation. 2008;3(3):134-6. Epub 2008 Nov 5.

FlexPred: a web-server for predicting residue positions involved in

conformational switches in proteins.

Kuznetsov IB(1), McDuffie M.

Author information:

(1)GenNY\*sis Center for Excellence in Cancer Genomics, Department of Epidemiology

and Biostatistics, One Discovery Drive, University at Albany, Rensselaer, NY

12144, USA. ikuznetsov@albany.edu

Conformational switches observed in the protein backbone play a key role in a

variety of fundamental biological activities. This paper describes a web-server

that implements a pattern recognition algorithm trained on the examples from the

Database of Macromolecular Movements to predict residue positions involved in

conformational switches. Prediction can be performed at an adjustable false

positive rate using a user-supplied protein sequence in FASTA format or a

structure in a Protein Data Bank (PDB) file. If a protein sequence is submitted,

then the web-server uses sequence-derived information only (such as evolutionary

conservation of residue positions). If a PDB file is submitted, then the

web-server uses sequence-derived information and residue solvent accessibility

calculated from this file.AVAILABILITY: FlexPred is publicly available at

http://flexpred.rit.albany.edu.

PMCID: PMC2639688

PMID: 19238251

2530. Comput Syst Bioinformatics Conf. 2008;7:195-202.

Combining sequence and structural profiles for protein solvent accessibility

prediction.

Bondugula R(1), Xu D.

Author information:

(1)Digital Biology Laboratory, 110 C.S. Bond Life Sciences Center, University of

Missouri, Columbia, MO 65211, USA. raj@bioanalysis.org

Solvent accessibility is an important structural feature for a protein. We

propose a new method for solvent accessibility prediction that uses known

structure and sequence information more efficiently. We first estimate the

relative solvent accessibility of the query protein using fuzzy mean operator

from the solvent accessibilities of known structure fragments that have similar

sequences to the query protein. We then integrate the estimated solvent

accessibility and the position specific scoring matrix of the query protein using

a neural network. We tested our method on a large data set consisting of 3386

non-redundant proteins. The comparison with other methods show slightly improved

prediction accuracies with our method. The resulting system does need not be

re-trained when new data is available. We incorporated our method into the MUPRED

system, which is available as a web server at http://digbio.missouri.edu/mupred.

PMCID: PMC2791713

PMID: 19642280 [Indexed for MEDLINE]

2531. Comput Syst Bioinformatics Conf. 2008;7:157-68.

Matching of structural motifs using hashing on residue labels and geometric

filtering for protein function prediction.

Moll M(1), Kavraki LE.

Author information:

(1)Department of Computer Science, Rice University, Houston, TX 77005, USA.

mmoll@cs.rice.edu

There is an increasing number of proteins with known structure but unknown

function. Determining their function would have a significant impact on

understanding diseases and designing new therapeutics. However, experimental

protein function determination is expensive and very time-consuming.

Computational methods can facilitate function determination by identifying

proteins that have high structural and chemical similarity. Our focus is on

methods that determine binding site similarity. Although several such methods

exist, it still remains a challenging problem to quickly find all

functionally-related matches for structural motifs in large data sets with high

specificity. In this context, a structural motif is a set of 3D points annotated

with physicochemical information that characterize a molecular function. We

propose a new method called LabelHash that creates hash tables of n-tuples of

residues for a set of targets. Using these hash tables, we can quickly look up

partial matches to a motif and expand those matches to complete matches. We show

that by applying only very mild geometric constraints we can find statistically

significant matches with extremely high specificity in very large data sets and

for very general structural motifs. We demonstrate that our method requires a

reasonable amount of storage when employing a simple geometric filter and further

improves on the specificity of our previous work while maintaining very high

sensitivity. Our algorithm is evaluated on 20 homolog classes and a non-redundant

version of the Protein Data Bank as our background data set. We use cluster

analysis to analyze why certain classes of homologs are more difficult to

classify than others. The LabelHash algorithm is implemented on a web server at

http://kavrakilab.org/labelhash/.

PMID: 19642277 [Indexed for MEDLINE]

2532. Comput Syst Bioinformatics Conf. 2008;7:121-32.

Predicting flexible length linear B-cell epitopes.

El-Manzalawy Y(1), Dobbs D, Honavar V.

Author information:

(1)Artificial Intelligence Laboratory, Iowa State University, Ames, IA 50010,

USA. yasser@iastate.edu

Identifying B-cell epitopes play an important role in vaccine design,

immunodiagnostic tests, and antibody production. Therefore, computational tools

for reliably predicting B-cell epitopes are highly desirable. We explore two

machine learning approaches for predicting flexible length linear B-cell

epitopes. The first approach utilizes four sequence kernels for determining a

similarity score between any arbitrary pair of variable length sequences. The

second approach utilizes four different methods of mapping a variable length

sequence into a fixed length feature vector. Based on our empirical comparisons,

we propose FBCPred, a novel method for predicting flexible length linear B-cell

epitopes using the subsequence kernel. Our results demonstrate that FBCPred

significantly outperforms all other classifiers evaluated in this study. An

implementation of FBCPred and the datasets used in this study are publicly

available through our linear B-cell epitope prediction server, BCPREDS, at:

http://ailab.cs.iastate.edu/bcpreds/.

PMCID: PMC3400678

PMID: 19642274 [Indexed for MEDLINE]

2533. Comput Syst Bioinformatics Conf. 2008;7:109-20.

Feedback algorithm and web-server for protein structure alignment.

Zhao Z(1), Fu B, Alanis FJ, Summa CM.

Author information:

(1)Department of Computer Science, University of New Orleans, New Orleans, LA

70148, USA. zzha2@cs.uno.edu

We have developed a feedback algorithm for protein structure alignment between

two protein backbones. A web portal implementing this method has been constructed

and is freely available for use at http://fpsa.cs.uno.edu/ with a mirror site at

http://fpsa.cs.panam.edu/FPSA/. We compare our algorithm with three other,

commonly used methods: CE, DaliLite and SSM. The results show that in most cases

our algorithm outputs a larger number of aligned positions when the (Calpha) RMSD

is comparable. Also, in many cases where the number of aligned positions is

larger or comparable, our learning method is able to achieve a smaller (Calpha)

RMSD than the other methods tested. This trend of larger number of aligned

positions and smaller (Calpha) RMSD is observed more frequently in cases where

the similarity between protein structures is weak.

PMID: 19642273 [Indexed for MEDLINE]

2534. Conf Proc IEEE Eng Med Biol Soc. 2008;2008:775-8. doi:

10.1109/IEMBS.2008.4649267.

Consultation virtual collaborative environment for 3D medicine.

Krsek P(1), Spanel M, Svub M, Stancl V, Siler O, Sára V.

Author information:

(1)Faculty of Information Technology, Brno University of Technology, Bozetechova

2, 612 66, Czech republic. krsek@fit.vutbr.cz

This article focuses on the problems of consultation virtual collaborative

environment, which is designed to support 3D medical applications. This system

allows loading CT/MR data from PACS system, segmentation and 3D models of

tissues. It allows distant 3D consultations of the data between technicians and

surgeons. System is designed as three-layer client-server architecture.

Communication between clients and server is done via HTTP/HTTPS protocol. Results

and tests have confirmed, that today's standard network latency and dataflow do

not affect the usability of our system.

DOI: 10.1109/IEMBS.2008.4649267

PMID: 19162770 [Indexed for MEDLINE]

2535. Genome Biol. 2008;9(11):R159. doi: 10.1186/gb-2008-9-11-r159. Epub 2008 Nov 12.

ArrayPlex: distributed, interactive and programmatic access to genome sequence,

annotation, ontology, and analytical toolsets.

Killion PJ(1), Iyer VR.

Author information:

(1)Center for Systems and Synthetic Biology, Institute for Cellular and Molecular

Biology, University of Texas at Austin, 1 University Station A4800, Austin, Texas

78712, USA. pkillion@mail.utexas.edu

ArrayPlex is a software package that centrally provides a large number of

flexible toolsets useful for functional genomics, including microarray data

storage, quality assessments, data visualization, gene annotation retrieval,

statistical tests, genomic sequence retrieval and motif analysis. It uses a

client-server architecture based on open source components, provides graphical,

command-line, and programmatic access to all needed resources, and is extensible

by virtue of a documented application programming interface. ArrayPlex is

available at http://sourceforge.net/projects/arrayplex/.

DOI: 10.1186/gb-2008-9-11-r159

PMCID: PMC2614491

PMID: 19014503 [Indexed for MEDLINE]

2536. Genome Biol. 2008;9 Suppl 2:S9. doi: 10.1186/gb-2008-9-s2-s9. Epub 2008 Sep 1.

Concept recognition for extracting protein interaction relations from biomedical

text.

Baumgartner WA Jr(1), Lu Z, Johnson HL, Caporaso JG, Paquette J, Lindemann A,

White EK, Medvedeva O, Cohen KB, Hunter L.

Author information:

(1)Center for Computational Pharmacology, University of Colorado School of

Medicine, Aurora, Colorado 80045, USA.

BACKGROUND: Reliable information extraction applications have been a long sought

goal of the biomedical text mining community, a goal that if reached would

provide valuable tools to benchside biologists in their increasingly difficult

task of assimilating the knowledge contained in the biomedical literature. We

present an integrated approach to concept recognition in biomedical text. Concept

recognition provides key information that has been largely missing from previous

biomedical information extraction efforts, namely direct links to well defined

knowledge resources that explicitly cement the concept's semantics. The

BioCreative II tasks discussed in this special issue have provided a unique

opportunity to demonstrate the effectiveness of concept recognition in the field

of biomedical language processing.

RESULTS: Through the modular construction of a protein interaction relation

extraction system, we present several use cases of concept recognition in

biomedical text, and relate these use cases to potential uses by the benchside

biologist.

CONCLUSION: Current information extraction technologies are approaching

performance standards at which concept recognition can begin to deliver high

quality data to the benchside biologist. Our system is available as part of the

BioCreative Meta-Server project and on the internet

http://bionlp.sourceforge.net.

DOI: 10.1186/gb-2008-9-s2-s9

PMCID: PMC2559993

PMID: 18834500 [Indexed for MEDLINE]

2537. Genome Biol. 2008;9 Suppl 2:S6. doi: 10.1186/gb-2008-9-s2-s6. Epub 2008 Sep 1.

Introducing meta-services for biomedical information extraction.

Leitner F(1), Krallinger M, Rodriguez-Penagos C, Hakenberg J, Plake C, Kuo CJ,

Hsu CN, Tsai RT, Hung HC, Lau WW, Johnson CA, Saetre R, Yoshida K, Chen YH, Kim

S, Shin SY, Zhang BT, Baumgartner WA Jr, Hunter L, Haddow B, Matthews M, Wang X,

Ruch P, Ehrler F, Ozgür A, Erkan G, Radev DR, Krauthammer M, Luong T, Hoffmann R,

Sander C, Valencia A.

Author information:

(1)Structural Biology and Biocomputing Programme, Spanish National Cancer

Research Centre (CNIO), Madrid, Spain. fleitner@cnio.es

We introduce the first meta-service for information extraction in molecular

biology, the BioCreative MetaServer (BCMS; http://bcms.bioinfo.cnio.es/). This

prototype platform is a joint effort of 13 research groups and provides

automatically generated annotations for PubMed/Medline abstracts. Annotation

types cover gene names, gene IDs, species, and protein-protein interactions. The

annotations are distributed by the meta-server in both human and machine readable

formats (HTML/XML). This service is intended to be used by biomedical researchers

and database annotators, and in biomedical language processing. The platform

allows direct comparison, unified access, and result aggregation of the

annotations.

DOI: 10.1186/gb-2008-9-s2-s6

PMCID: PMC2559990

PMID: 18834497 [Indexed for MEDLINE]

2538. Genome Biol. 2008;9(8):R128. doi: 10.1186/gb-2008-9-8-r128. Epub 2008 Aug 15.

MotifCluster: an interactive online tool for clustering and visualizing sequences

using shared motifs.

Hamady M(1), Widmann J, Copley SD, Knight R.

Author information:

(1)Department of Computer Science, University of Colorado, Boulder, CO 80309,

USA.

MotifCluster finds related motifs in a set of sequences, and clusters the

sequences into families using the motifs they contain. MotifCluster, at

http://bmf.colorado.edu/motifcluster, lets users test whether proteins are

related, cluster sequences by shared conserved motifs, and visualize motifs

mapped onto trees, sequences and three-dimensional structures. We demonstrate

MotifCluster's accuracy using gold-standard protein superfamilies; using

recommended settings, families were assigned to the correct superfamilies with

0.17% false positive and no false negative assignments.

DOI: 10.1186/gb-2008-9-8-r128

PMCID: PMC2575518

PMID: 18706079 [Indexed for MEDLINE]

2539. Genome Biol. 2008;9(6):R96. doi: 10.1186/gb-2008-9-6-r96. Epub 2008 Jun 12.

Anni 2.0: a multipurpose text-mining tool for the life sciences.

Jelier R(1), Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA.

Author information:

(1)Department of Medical Informatics, Erasmus MC University Medical Center, Dr,

Molewaterplein, Rotterdam, 3015 GE, The Netherlands. r.jelier@erasmusmc.nl.

Anni 2.0 is an online tool (http://biosemantics.org/anni/) to aid the biomedical

researcher with a broad range of information needs. Anni provides an

ontology-based interface to MEDLINE and retrieves documents and associations for

several classes of biomedical concepts, including genes, drugs and diseases, with

established text-mining technology. In this article we illustrate Anni's

usability by applying the tool to two use cases: interpretation of a set of

differentially expressed genes, and literature-based knowledge discovery.

DOI: 10.1186/gb-2008-9-6-r96

PMCID: PMC2481428

PMID: 18549479 [Indexed for MEDLINE]

2540. In Silico Biol. 2008;8(3-4):223-34.

PredictBias: a server for the identification of genomic and pathogenicity islands

in prokaryotes.

Pundhir S(1), Vijayvargiya H, Kumar A.

Author information:

(1)Bioinformatics sub-centre, School of Biotechnology, Devi Ahilya University,

Khandwa Road, Indore, India.

Pathogenicity Islands (PAIs) are the sub-sets of Genomic Islands (GIs) that are

acquired by horizontal gene transfer (HGT) and are generally shown to have a

significant deviation in G+C, dinucleotide or codon frequency from core genome.

Major approaches used for PAI identification are based on composition bias and/or

similarity with known PAIs. These approaches either limit the search to GIs or to

regions similar to previously annotated PAIs. PredictBias is a web application

for the identification of genomic and pathogenicity islands in prokaryotes based

on composition bias, presence of insertion elements, proximity with

virulence-associated genes and absence in related non-pathogenic species. A

profile database of virulence factors (VFPD) has been developed using 213 protein

families associated to virulence retrieved from Pfam and PRINTS database.

PredictBias performs a RPSBLAST search for regions with significant composition

bias against VFPD. If a region encodes for at least one protein related to

virulence then it is marked as potential PAI (biased-composition) otherwise as

GI. Regions involved in virulence but having unsuspicious composition bias due to

ancient HGT are identified by scanning genome segments (8 ORFs) with more than

four significant hits to VFPD and are marked as potential PAI

(unbiased-composition). The relative absence of potential PAIs in related

non-pathogenic species can be investigated using 'compare genome feature' of

PredictBias that further aids in validating the results and defining boundaries

for PAIs. Performance measure analysis showed that the output of PredictBias is

in agreement with the known findings. PredictBias is available at

www.davvbiotech.res.in/PredictBias.

PMID: 19032158 [Indexed for MEDLINE]

2541. In Silico Biol. 2008;8(2):129-40.

A machine learning based method for the prediction of secretory proteins using

amino acid composition, their order and similarity-search.

Garg A(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector 39A, Chandigarh, India.

Most of the prediction methods for secretory proteins require the presence of a

correct N-terminal end of the preprotein for correct classification. As large

scale genome sequencing projects sometimes assign the 5'-end of genes

incorrectly, many proteins are encoded without the correct N-terminus leading to

incorrect prediction. In this study, a systematic attempt has been made to

predict secretory proteins irrespective of presence or absence of N-terminal

signal peptides (also known as classical and non-classical secreted proteins

respectively), using machine-learning techniques; artificial neural network (ANN)

and support vector machine (SVM). We trained and tested our methods on a dataset

of 3321 secretory and 3654 non-secretory mammalian proteins using five-fold

cross-validation technique. First, ANN-based modules have been developed for

predicting secretory proteins using 33 physico-chemical properties, amino acid

composition and dipeptide composition and achieved accuracies of 73.1%, 76.1% and

77.1%, respectively. Similarly, SVM-based modules using 33 physico-chemical

properties, amino acid, and dipeptide composition have been able to achieve

accuracies of 77.4%, 79.4% and 79.9%, respectively. In addition, BLAST and

PSI-BLAST modules designed for predicting secretory proteins based on similarity

search achieved 23.4% and 26.9% accuracy, respectively. Finally, we developed a

hybrid-approach by integrating amino acid and dipeptide composition based SVM

modules and PSI-BLAST module that increased the accuracy to 83.2%, which is

significantly better than individual modules. We also achieved high sensitivity

of 60.4% with low value of 5% false positive predictions using hybrid module. A

web server SRTpred has been developed based on above study for predicting

classical and non-classical secreted proteins from whole sequence of mammalian

proteins, which is available from http://www.imtech.res.in/raghava/srtpred/.

PMID: 18928201 [Indexed for MEDLINE]

2542. In Silico Biol. 2008;8(2):121-8.

COPid: composition based protein identification.

Kumar M(1), Thakur V, Raghava GP.

Author information:

(1)Institute ofMicrobial Technology, Sector 39-A, Chandigarh, India.

In the past, a large number of methods have been developed for predicting various

characteristics of a protein from its composition. In order to exploit the full

potential of protein composition, we developed the web-server COPid to assist the

researchers in annotating the function of a protein from its composition using

whole or part of the protein. COPid has three modules called search, composition

and analysis. The search module allows searching of protein sequences in six

different databases. Search results list database proteins in ascending order of

Euclidian distance or descending order of compositional similarity with the query

sequence. The composition module allows calculation of the composition of a

sequence and average composition of a group of sequences. The composition module

also allows computing composition of various types of amino acids (e.g. charge,

polar, hydrophobic residues). The analysis module provides the following options;

i) comparing composition of two classes of proteins, ii) creating a phylogenetic

tree based on the composition and iii) generating input patterns for machine

learning techniques. We have evaluated the performance of composition-based (or

alignment-free) similarity search in the subcellular localization of proteins. It

was found that the alignment free method performs reasonably well in predicting

certain classes of proteins. The COPid web-server is available at

http://www.imtech.res.in/raghava/copid/.

PMID: 18928200 [Indexed for MEDLINE]

2543. Int J Comput Biol Drug Des. 2008;1(3):302-12.

Implementing bioinformatic workflows within the bioextract server.

Lushbough CM(1), Bergman MK, Lawrence CJ, Jennewein D, Brendel V.

Author information:

(1)Department of Computer Science, University of South Dakota, Vermillion, SD

57069, USA. Carol.Lushbough@usd.edu

Computational workflows in bioinformatics are becoming increasingly important in

the achievement of scientific advances. These workflows typically require the

integrated use of multiple, distributed data sources and analytic tools. The

BioExtract Server (http://bioextract.org) is a distributed service designed to

provide researchers with the web ability to query multiple data sources, save

results as searchable data sets, and execute analytic tools. As the researcher

works with the system, their tasks are saved in the background. At any time these

steps can be saved as a workflow that can then be executed again and/or modified

later.

PMID: 20054995 [Indexed for MEDLINE]

2544. Int J Occup Saf Ergon. 2008;14(4):379-86.

Occupational exposure to mechanical vibration: the Italian vibration database for

risk assessment.

Nataletti P(1), Marchetti E, Lunghi A, Pinto I, Stacchini N, Santini F.

Author information:

(1)Department of Occupational Hygiene, National Institute of Occupational

Prevention and Safety (ISPESL), Rome, Italy. pietro.nataletti@ispesl.it

The Italian vibration database is presented. It is hosted by a web server at the

National Institute of Occupational Prevention and Safety (ISPESL) in Rome, Italy

(http: / /www.ispesl.it/vibrationdatabase). It supports in risk assessment

employers who have to comply with Legislative Decree 187/05, now replaced by

Legislative Decree 81/08, which transposes into law Vibration Directive

2002/44/EC. The database currently contains measurements and EC-declared values

from over 980 hand-held power tools (such as pneumatic and electric hammers,

chainsaws, grinders, drills, sanders and saws) and from over 420 vehicles (such

as buses, forklifts and wheel tractors). The database is continuously updated as

soon as new experimental and declared data are acquired.

DOI: 10.1080/10803548.2008.11076775

PMID: 19080042 [Indexed for MEDLINE]

2545. J Biomol NMR. 2008 Jan;40(1):31-48. Epub 2007 Nov 6.

Application of the random coil index to studying protein flexibility.

Berjanskii MV(1), Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB, Canada.

Protein flexibility lies at the heart of many protein-ligand binding events and

enzymatic activities. However, the experimental measurement of protein motions is

often difficult, tedious and error-prone. As a result, there is a considerable

interest in developing simpler and faster ways of quantifying protein

flexibility. Recently, we described a method, called Random Coil Index (RCI),

which appears to be able to quantitatively estimate model-free order parameters

and flexibility in protein structural ensembles using only backbone chemical

shifts. Because of its potential utility, we have undertaken a more detailed

investigation of the RCI method in an attempt to ascertain its underlying

principles, its general utility, its sensitivity to chemical shift errors, its

sensitivity to data completeness, its applicability to other proteins, and its

general strengths and weaknesses. Overall, we find that the RCI method is very

robust and that it represents a useful addition to traditional methods of

studying protein flexibility. We have implemented many of the findings and

refinements reported here into a web server that allows facile, automated

predictions of model-free order parameters, MD RMSF and NMR RMSD values directly

from backbone 1H, 13C and 15N chemical shift assignments. The server is available

at http://wishart.biology.ualberta.ca/rci.

DOI: 10.1007/s10858-007-9208-0

PMID: 17985196 [Indexed for MEDLINE]

2546. J Clin Lab Anal. 2008;22(1):77-85. doi: 10.1002/jcla.20214.

Mass spectroscopic characteristics of low molecular weight proteins extracted

from calcium oxalate stones: preliminary study.

Chen WC(1), Lai CC, Tsai Y, Lin WY, Tsai FJ.

Author information:

(1)Department of Urology, China Medical University Hospital, Taichung, Taiwan.

wgchen@mail.cmu.edu.tw

Erratum in

J Clin Lab Anal. 2008;22(2):151. Lai, Chein-Cheng [corrected to Lai, Chien-Chen];

Tsai, Yu-Hsin [corrected to Tsai, Yuhsin].

It is believed that boundary compositions of matrix proteins might play a role in

stone formation; however, few proteomic studies concerning matrix proteins in

urinary stones have been conducted. In this study, we extracted low molecular

weight proteins from calcium oxalate stones and measured their characteristic

patterns by mass spectroscopy. A total of 10 stones were surgically removed from

patients with urolithiasis. Proteins were extracted from the stones and

identified by one-dimensional electrophoresis (sodium dodecyl sulfate buffer

[SDS]-polyacrylamide gel electrophoresis [SDS-PAGE]). After in-gel digest,

samples were analyzed by the surface-enhanced laser desorption ionization-time of

flight (SELDI-TOF) technique. The peptide sequences were analyzed from the data

of mass spectroscopy. Proteins were identified from Database Search (SwissProt

Protein Database; Swiss Institute of Bioinformatics; http://www.expasy.org/sprot)

on a MASCOT server (Matrix Science Ltd.; http://www.matrixscience.com). A total

of three bands of proteins (27, 18, and 14 kDa) were identified from SDS-PAGE in

each stone sample. A database search (SwissProt) on a MASCOT server revealed that

the most frequently seen proteins from band 1 (27 kDa) were leukocyte elastase

precursor, cathepsin G precursor, azurocidin precursor, and myeloblastin

precursor (EC 3.4.21.76) (leukocyte proteinase 3); band 2 (18 kDa) comprised

calgranulin B, eosinophil cationic protein precursor, and lysozyme C precursor;

band 3 (14 kDa) showed neutrophil defensin 3 precursor, calgranulin A,

calgranulin C, and histone H4. The modifications and deamidations found from the

mass pattern of these proteins may provide information for the study of matrix

proteins. Various lower molecular weight proteins can be extracted from calcium

oxalate stones. The characteristic patterns and their functions of those proteins

should be further tested to investigate their roles in stone formation.

(c) 2008 Wiley-Liss, Inc.

DOI: 10.1002/jcla.20214

PMID: 18200570 [Indexed for MEDLINE]

2547. J Mol Model. 2008 Jan;14(1):69-76. Epub 2007 Nov 8.

AutoMotif Server for prediction of phosphorylation sites in proteins using

support vector machine: 2007 update.

Plewczynski D(1), Tkacz A, Wyrwicz LS, Rychlewski L, Ginalski K.

Author information:

(1)Interdisciplinary Centre for Mathematical and Computational Modeling,

University of Warsaw, Pawinskiego 5a, 02-106, Warsaw, Poland.

D.Plewczynski@icm.edu.pl

We present here the recent update of AutoMotif Server (AMS 2.0) that predicts

post-translational modification sites in protein sequences. The support vector

machine (SVM) algorithm was trained on data gathered in 2007 from various sets of

proteins containing experimentally verified chemical modifications of proteins.

Short sequence segments around a modification site were dissected from a parent

protein, and represented in the training set as binary or profile vectors. The

updated efficiency of the SVM classification for each type of modification and

the predictive power of both representations were estimated using leave-one-out

tests for model of general phosphorylation and for modifications catalyzed by

several specific protein kinases. The accuracy of the method was improved in

comparison to the previous version of the service (Plewczynski et al., "AutoMotif

server: prediction of single residue post-translational modifications in

proteins", Bioinformatics 21: 2525-7, 2005). The precision of the updated version

reached over 90% for selected types of phosphorylation and was optimized in trade

of lower recall value of the classification model. The AutoMotif Server version

2007 is freely available at http://ams2.bioinfo.pl/ . Additionally, the reference

dataset for optimization of prediction of phosphorylation sites, collected from

the UniProtKB was also provided and can be accessed at

http://ams2.bioinfo.pl/data/ .

DOI: 10.1007/s00894-007-0250-3

PMID: 17994256 [Indexed for MEDLINE]

2548. Methods Mol Biol. 2008;420:45-59. doi: 10.1007/978-1-59745-583-1\_3.

FlyBase : a database for the Drosophila research community.

Drysdale R(1); FlyBase Consortium.

Author information:

(1)Department of Genetics, University of Cambridge, Cambridge, UK.

FlyBase ( http://flybase.org ) is the primary database of integrated genetic and

genomic data about the Drosophilidae, of which Drosophila melanogaster is the

most extensively studied species. Information in FlyBase originates from a

variety of sources ranging from large-scale genome projects to the primary

research literature. Data-types include sequence-level gene models, molecular

classification of gene product functions, mutant phenotypes, mutant lesions and

chromosome aberrations, gene expression patterns, transgene insertions, and

anatomical images. Query tools allow interrogation of FlyBase through DNA or

protein sequence, by gene or mutant name, or through terms from the several

ontologies used to capture functional, phenotypic, and anatomical data. Links

between FlyBase and external databases provide extensive opportunity for

extending exploration into other model organism databases and resources of

biological and molecular information. This review will introduce the FlyBase web

server and query tools.

DOI: 10.1007/978-1-59745-583-1\_3

PMID: 18641940 [Indexed for MEDLINE]

2549. Methods Mol Biol. 2008;446:281-92. doi: 10.1007/978-1-60327-084-7\_20.

Analysis of O-glycosylation.

Calvete JJ(1), Sanz L.

Author information:

(1)Instituto de Biomedicina de Valencia, CSIC, Valencia, Spain.

Secreted as well as membrane-associated eukaryotic proteins are most commonly

glycosylated. Saccharides are attached to proteins mainly through N-and

O-glycosydic bonds or as part of the glycosylphosphatidyinositol-membrane anchor.

In contrast to N-glycosylation, which involves the co-translational transfer in

the endoplasmic reticulum (ER) of the glycan portion of

Glc3Man9GlcNAc2-PP-dolichol to suitable Asn residues on nascent polypeptides,

O-glycosylation begins with the addition of a single monosaccharide. Contrary to

N-glycosylation, which involves an asparagine residue in the sequon

Asn-Xaa-Thr/Ser (Xaa can be any amino acid except Pro, and it is rarely Cys), no

particular sequence motif has been described for O-glycosylation. This may

reflect the fact that: (1) the specificity of the UDP-GalNAc:polypeptide

N-acetylgalactosaminyltransferase is presently unknown; and (2) seems to be

modulated by sequence context, secondary structure, and surface accessibility

(1). An internet server, accessible at

http://www.cbs.dtu.dk/netOglyc/cbsnetOglyc.html , produces neural network

predictions of mucin-type GalNAc O-glycosylation sites in mammalian proteins

based on 299 known and verified mucin-type O-glycosylation sites. The sequence

context of glycosylated threonines was found to differ from that of serine, and

the sites were found to cluster. Nonclustered sites had a sequence context

different from that of clustered sites, and charged residues were disfavored at

position -1 and +3.

DOI: 10.1007/978-1-60327-084-7\_20

PMID: 18373265 [Indexed for MEDLINE]

2550. Nat Protoc. 2008;3(10):1578-88. doi: 10.1038/nprot.2008.97.

Using RSAT to scan genome sequences for transcription factor binding sites and

cis-regulatory modules.

Turatsinze JV(1), Thomas-Chollier M, Defrance M, van Helden J.

Author information:

(1)Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université

Libre de Bruxelles CP 263, Campus Plaine, Boulevard du Triomphe, Bruxelles,

Belgium.

This protocol shows how to detect putative cis-regulatory elements and regions

enriched in such elements with the regulatory sequence analysis tools (RSAT) web

server (http://rsat.ulb.ac.be/rsat/). The approach applies to known transcription

factors, whose binding specificity is represented by position-specific scoring

matrices, using the program matrix-scan. The detection of individual binding

sites is known to return many false predictions. However, results can be strongly

improved by estimating P value, and by searching for combinations of sites

(homotypic and heterotypic models). We illustrate the detection of sites and

enriched regions with a study case, the upstream sequence of the Drosophila

melanogaster gene even-skipped. This protocol is also tested on random control

sequences to evaluate the reliability of the predictions. Each task requires a

few minutes of computation time on the server. The complete protocol can be

executed in about one hour.

DOI: 10.1038/nprot.2008.97

PMID: 18802439 [Indexed for MEDLINE]

2551. Nat Protoc. 2008;3(2):153-62. doi: 10.1038/nprot.2007.494.

Cell-PLoc: a package of Web servers for predicting subcellular localization of

proteins in various organisms.

Chou KC(1), Shen HB.

Author information:

(1)Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego,

California 92130, USA. kcchou@gordonlifescience.org

Information on subcellular localization of proteins is important to molecular

cell biology, proteomics, system biology and drug discovery. To provide the vast

majority of experimental scientists with a user-friendly tool in these areas, we

present a package of Web servers developed recently by hybridizing the 'higher

level' approach with the ab initio approach. The package is called Cell-PLoc and

contains the following six predictors: Euk-mPLoc, Hum-mPLoc, Plant-PLoc,

Gpos-PLoc, Gneg-PLoc and Virus-PLoc, specialized for eukaryotic, human, plant,

Gram-positive bacterial, Gram-negative bacterial and viral proteins,

respectively. Using these Web servers, one can easily get the desired prediction

results with a high expected accuracy, as demonstrated by a series of

cross-validation tests on the benchmark data sets that covered up to 22

subcellular location sites and in which none of the proteins included had > or

=25% sequence identity to any other protein in the same subcellular-location

subset. Some of these Web servers can be particularly used to deal with multiplex

proteins as well, which may simultaneously exist at, or move between, two or more

different subcellular locations. Proteins with multiple locations or dynamic

features of this kind are particularly interesting, because they may have some

special biological functions intriguing to investigators in both basic research

and drug discovery. This protocol is a step-by-step guide on how to use the

Web-server predictors in the Cell-PLoc package. The computational time for each

prediction is less than 5 s in most cases. The Cell-PLoc package is freely

accessible at http://chou.med.harvard.edu/bioinf/Cell-PLoc.

DOI: 10.1038/nprot.2007.494

PMID: 18274516 [Indexed for MEDLINE]

2552. Nucleic Acids Res. 2008 Jan;36(Database issue):D196-201. Epub 2007 Dec 23.

MIPS: analysis and annotation of genome information in 2007.

Mewes HW(1), Dietmann S, Frishman D, Gregory R, Mannhaupt G, Mayer KF,

Münsterkötter M, Ruepp A, Spannagl M, Stümpflen V, Rattei T.

Author information:

(1)Institute for Bioinformatics (MIPS), German Research Center for Environmental

Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany.

The Munich Information Center for Protein Sequences (MIPS-GSF, Neuherberg,

Germany) combines automatic processing of large amounts of sequences with manual

annotation of selected model genomes. Due to the massive growth of the available

data, the depth of annotation varies widely between independent databases. Also,

the criteria for the transfer of information from known to orthologous sequences

are diverse. To cope with the task of global in-depth genome annotation has

become unfeasible. Therefore, our efforts are dedicated to three levels of

annotation: (i) the curation of selected genomes, in particular from fungal and

plant taxa (e.g. CYGD, MNCDB, MatDB), (ii) the comprehensive, consistent,

automatic annotation employing exhaustive methods for the computation of sequence

similarities and sequence-related attributes as well as the classification of

individual sequences (SIMAP, PEDANT and FunCat) and (iii) the compilation of

manually curated databases for protein interactions based on scrutinized

information from the literature to serve as an accepted set of reliable annotated

interaction data (MPACT, MPPI, CORUM). All databases and tools described as well

as the detailed descriptions of our projects can be accessed through the MIPS web

server (http://mips.gsf.de).

DOI: 10.1093/nar/gkm980

PMCID: PMC2238900

PMID: 18158298 [Indexed for MEDLINE]

2553. Nucleic Acids Res. 2008 Jan;36(Database issue):D959-65. Epub 2007 Dec 6.

PlantGDB: a resource for comparative plant genomics.

Duvick J(1), Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough

C, Brendel V.

Author information:

(1)Department of Genetics, Development and Cell Biology, Iowa State University,

Ames, IA 50011, USA.

PlantGDB (http://www.plantgdb.org/) is a genomics database encompassing sequence

data for green plants (Viridiplantae). PlantGDB provides annotated transcript

assemblies for >100 plant species, with transcripts mapped to their cognate

genomic context where available, integrated with a variety of sequence analysis

tools and web services. For 14 plant species with emerging or complete genome

sequence, PlantGDB's genome browsers (xGDB) serve as a graphical interface for

viewing, evaluating and annotating transcript and protein alignments to

chromosome or bacterial artificial chromosome (BAC)-based genome assemblies.

Annotation is facilitated by the integrated yrGATE module for community curation

of gene models. Novel web services at PlantGDB include Tracembler, an iterative

alignment tool that generates contigs from GenBank trace file data and BioExtract

Server, a web-based server for executing custom sequence analysis workflows.

PlantGDB also hosts a plant genomics research outreach portal (PGROP) that

facilitates access to a large number of resources for research and training.

DOI: 10.1093/nar/gkm1041

PMCID: PMC2238959

PMID: 18063570 [Indexed for MEDLINE]

2554. Nucleic Acids Res. 2008 Jan;36(Database issue):D469-74. Epub 2007 Nov 21.

GenoList: an integrated environment for comparative analysis of microbial

genomes.

Lechat P(1), Hummel L, Rousseau S, Moszer I.

Author information:

(1)Plate-forme Intégration et Analyse Génomiques, Pasteur Génopole Ile-de-France,

Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France.

The multitude of bacterial genome sequences being determined has generated new

requirements regarding the development of databases and graphical interfaces:

these are needed to organize and retrieve biological information from the

comparison of large sets of genomes. GenoList

(http://genolist.pasteur.fr/GenoList) is an integrated environment dedicated to

querying and analyzing genome data from bacterial species. GenoList inherits from

the SubtiList database and web server, the reference data resource for the

Bacillus subtilis genome. The data model was extended to hold information about

relationships between genomes (e.g. protein families). The web user interface was

designed to primarily take into account biologists' needs and modes of operation.

Along with standard query and browsing capabilities, comparative genomics

facilities are available, including subtractive proteome analysis. One key

feature is the integration of the many tools accessible in the environment. As an

example, it is straightforward to identify the genes that are specific to a group

of bacteria, export them as a tab-separated list, get their protein sequences and

run a multiple alignment on a subset of these sequences.

DOI: 10.1093/nar/gkm1042

PMCID: PMC2238853

PMID: 18032431 [Indexed for MEDLINE]

2555. Nucleic Acids Res. 2008 Jan;36(Database issue):D651-5. Epub 2007 Nov 13.

DIMA 2.0--predicted and known domain interactions.

Pagel P(1), Oesterheld M, Tovstukhina O, Strack N, Stümpflen V, Frishman D.

Author information:

(1)Lehrstuhl für Genomorientierte Bioinformatik, Wissenschaftszentrum

Weihenstephan, Technische Universität München, Am Forum 1, D-85350 Freising,

Germany. p.pagel@wzw.tum.de

DIMA-the domain interaction map has evolved from a simple web server for domain

phylogenetic profiling into an integrative prediction resource combining both

experimental data on domain-domain interactions and predictions from two

different algorithms. With this update, DIMA obtains greatly improved coverage at

the level of genomes and domains as well as with respect to available prediction

approaches. The domain phylogenetic profiling method now uses SIMAP as its

backend for exhaustive domain hit coverage: 7038 Pfam domains were profiled over

460 completely sequenced genomes. Domain pair exclusion predictions were produced

from 83 969 distinct protein-protein interactions obtained from IntAct resulting

in 21 513 domain pairs with significant domain pair exclusion algorithm scores.

Additional predictions applying the same algorithm to predicted protein

interactions from STRING yielded 2378 high-confidence pairs. Experimental data

comes from iPfam (3074) and 3did (3034 pairs), two databases identifying domain

contacts in solved protein structures. Taken together, these two resources

yielded 3653 distinct interacting domain pairs. DIMA is available at

http://mips.gsf.de/genre/proj/dima.

DOI: 10.1093/nar/gkm996

PMCID: PMC2238836

PMID: 17999995 [Indexed for MEDLINE]

2556. Nucleic Acids Res. 2008 Jan;36(Database issue):D768-72. Epub 2007 Nov 8.

The Zebrafish Information Network: the zebrafish model organism database provides

expanded support for genotypes and phenotypes.

Sprague J(1), Bayraktaroglu L, Bradford Y, Conlin T, Dunn N, Fashena D, Frazer K,

Haendel M, Howe DG, Knight J, Mani P, Moxon SA, Pich C, Ramachandran S, Schaper

K, Segerdell E, Shao X, Singer A, Song P, Sprunger B, Van Slyke CE, Westerfield

M.

Author information:

(1)The Zebrafish Information Network, 5291 University of Oregon, Eugene, OR

97403-5291, USA. judys@cs.uoregon.edu

The Zebrafish Information Network (ZFIN, http://zfin.org), the model organism

database for zebrafish, provides the central location for curated zebrafish

genetic, genomic and developmental data. Extensive data integration of mutant

phenotypes, genes, expression patterns, sequences, genetic markers, morpholinos,

map positions, publications and community resources facilitates the use of the

zebrafish as a model for studying gene function, development, behavior and

disease. Access to ZFIN data is provided via web-based query forms and through

bulk data files. ZFIN is the definitive source for zebrafish gene and allele

nomenclature, the zebrafish anatomical ontology (AO) and for zebrafish gene

ontology (GO) annotations. ZFIN plays an active role in the development of

cross-species ontologies such as the phenotypic quality ontology (PATO) and the

gene ontology (GO). Recent enhancements to ZFIN include (i) a new home page and

navigation bar, (ii) expanded support for genotypes and phenotypes, (iii)

comprehensive phenotype annotations based on anatomical, phenotypic quality and

gene ontologies, (iv) a BLAST server tightly integrated with the ZFIN database

via ZFIN-specific datasets, (v) a global site search and (vi) help with hands-on

resources.

DOI: 10.1093/nar/gkm956

PMCID: PMC2238839

PMID: 17991680 [Indexed for MEDLINE]

2557. Nucleic Acids Res. 2008 Jan;36(Database issue):D871-7. Epub 2007 Nov 7.

The Stanford Tissue Microarray Database.

Marinelli RJ(1), Montgomery K, Liu CL, Shah NH, Prapong W, Nitzberg M, Zachariah

ZK, Sherlock GJ, Natkunam Y, West RB, van de Rijn M, Brown PO, Ball CA.

Author information:

(1)Department of Biochemistry, Stanford University School of Medicine, Howard

Hughes Medical Institute, Stanford, CA, USA. bobm@stanford.edu

The Stanford Tissue Microarray Database (TMAD; http://tma.stanford.edu) is a

public resource for disseminating annotated tissue images and associated

expression data. Stanford University pathologists, researchers and their

collaborators worldwide use TMAD for designing, viewing, scoring and analyzing

their tissue microarrays. The use of tissue microarrays allows hundreds of human

tissue cores to be simultaneously probed by antibodies to detect protein

abundance (Immunohistochemistry; IHC), or by labeled nucleic acids (in situ

hybridization; ISH) to detect transcript abundance. TMAD archives

multi-wavelength fluorescence and bright-field images of tissue microarrays for

scoring and analysis. As of July 2007, TMAD contained 205 161 images archiving

349 distinct probes on 1488 tissue microarray slides. Of these, 31 306 images for

68 probes on 125 slides have been released to the public. To date, 12

publications have been based on these raw public data. TMAD incorporates the NCI

Thesaurus ontology for searching tissues in the cancer domain. Image processing

researchers can extract images and scores for training and testing classification

algorithms. The production server uses the Apache HTTP Server, Oracle Database

and Perl application code. Source code is available to interested researchers

under a no-cost license.

DOI: 10.1093/nar/gkm861

PMCID: PMC2238948

PMID: 17989087 [Indexed for MEDLINE]

2558. Nucleic Acids Res. 2008 Jan;36(Database issue):D719-23. Epub 2007 Oct 11.

Gallus GBrowse: a unified genomic database for the chicken.

Schmidt CJ(1), Romanov M, Ryder O, Magrini V, Hickenbotham M, Glasscock J,

McGrath S, Mardis E, Stein LD.

Author information:

(1)Department of Animal and Food Sciences, University of Delaware, Newark, DE

19706, USA. schmidtc@udel.edu

Gallus GBrowse (http://birdbase.net/cgi-bin/gbrowse/gallus/) provides online

access to genomic and other information about the chicken, Gallus gallus. The

information provided by this resource includes predicted genes and Gene Ontology

(GO) terms, links to Gallus In Situ Hybridization Analysis (GEISHA), Unigene and

Reactome, the genomic positions of chicken genetic markers, SNPs and microarray

probes, and mappings from turkey, condor and zebra finch DNA and EST sequences to

the chicken genome. We also provide a BLAT server

(http://birdbase.net/cgi-bin/webBlat) for matching user-provided sequences to the

chicken genome. These tools make the Gallus GBrowse server a valuable resource

for researchers seeking genomic information regarding the chicken and other avian

species.

DOI: 10.1093/nar/gkm783

PMCID: PMC2238981

PMID: 17933775 [Indexed for MEDLINE]

2559. Nucleic Acids Res. 2008 Jan;36(Database issue):D667-73. Epub 2007 Oct 11.

LigASite--a database of biologically relevant binding sites in proteins with

known apo-structures.

Dessailly BH(1), Lensink MF, Orengo CA, Wodak SJ.

Author information:

(1)Center for Structural Biology and Bioinformatics, Université Libre de

Bruxelles (U. L. B.), Bld du Triomphe - CP 263, 1050 Bruxelles, Belgium.

Better characterization of binding sites in proteins and the ability to

accurately predict their location and energetic properties are major challenges

which, if addressed, would have many valuable practical applications.

Unfortunately, reliable benchmark datasets of binding sites in proteins are still

sorely lacking. Here, we present LigASite ('LIGand Attachment SITE'), a

gold-standard dataset of binding sites in 550 proteins of known structures.

LigASite consists exclusively of biologically relevant binding sites in proteins

for which at least one apo- and one holo-structure are available. In defining the

binding sites for each protein, information from all holo-structures is combined,

considering in each case the quaternary structure defined by the PQS server.

LigASite is built using simple criteria and is automatically updated as new

structures become available in the PDB, thereby guaranteeing optimal data

coverage over time. Both a redundant and a culled non-redundant version of the

dataset is available at http://www.scmbb.ulb.ac.be/Users/benoit/LigASite. The

website interface allows users to search the dataset by PDB identifiers, ligand

identifiers, protein names or sequence, and to look for structural matches as

defined by the CATH homologous superfamilies. The datasets can be downloaded from

the website as Schema-validated XML files or comma-separated flat files.

DOI: 10.1093/nar/gkm839

PMCID: PMC2238865

PMID: 17933762 [Indexed for MEDLINE]

2560. Nucleic Acids Res. 2008 Jan;36(Database issue):D1034-40. Epub 2007 Oct 11.

GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics

and genetics data.

Jung S(1), Staton M, Lee T, Blenda A, Svancara R, Abbott A, Main D.

Author information:

(1)Department of Horticulture and Landscape Architecture, Washington State

University, Pullman, WA 99164-6414, USA. sook@bioinfo.wsu.edu

The Genome Database for Rosaceae (GDR) is a central repository of curated and

integrated genetics and genomics data of Rosaceae, an economically important

family which includes apple, cherry, peach, pear, raspberry, rose and strawberry.

GDR contains annotated databases of all publicly available Rosaceae ESTs, the

genetically anchored peach physical map, Rosaceae genetic maps and

comprehensively annotated markers and traits. The ESTs are assembled to produce

unigene sets of each genus and the entire Rosaceae. Other annotations include

putative function, microsatellites, open reading frames, single nucleotide

polymorphisms, gene ontology terms and anchored map position where applicable.

Most of the published Rosaceae genetic maps can be viewed and compared through

CMap, the comparative map viewer. The peach physical map can be viewed using

WebFPC/WebChrom, and also through our integrated GDR map viewer, which serves as

a portal to the combined genetic, transcriptome and physical mapping information.

ESTs, BACs, markers and traits can be queried by various categories and the

search result sites are linked to the mapping visualization tools. GDR also

provides online analysis tools such as a batch BLAST/FASTA server for the GDR

datasets, a sequence assembly server and microsatellite and primer detection

tools. GDR is available at http://www.rosaceae.org.

DOI: 10.1093/nar/gkm803

PMCID: PMC2238863

PMID: 17932055 [Indexed for MEDLINE]

2561. Nucleic Acids Res. 2008 Jan;36(Database issue):D409-13. Epub 2007 Oct 5.

coliSNP database server mapping nsSNPs on protein structures.

Kono H(1), Yuasa T, Nishiue S, Yura K.

Author information:

(1)Computational Biology Group, Quantum Beam Science Directorate, Japan Atomic

Energy Agency, 8-1 Umemidai, Kizugawa, Kyoto 619-0215, PRESTO, Japan.

kono.hidetoshi@jaea.go.jp

We have developed coliSNP, a database server

(http://yayoi.kansai.jaea.go.jp/colisnp) that maps non-synonymous single

nucleotide polymorphisms (nsSNPs) on the three-dimensional (3D) structure of

proteins. Once a week, the SNP data from the dbSNP database and the protein

structure data from the Protein Data Bank (PDB) are downloaded, and the

correspondence of the two data sets is automatically tabulated in the coliSNP

database. Given an amino acid sequence, protein name or PDB ID, the server will

immediately provide known nsSNP information, including the amino acid mutation

caused by the nsSNP, the solvent accessibility, the secondary structure and the

flanking residues of the mutated residue in a single page. The position of the

nsSNP within the amino acid sequence and on the 3D structure of the protein can

also be observed. The database provides key information with which to judge

whether an observed nsSNP critically affects protein function and/or stability.

As far as we know, this is the only web-based nsSNP database that automatically

compiles SNP and protein information in a concise manner.

DOI: 10.1093/nar/gkm801

PMCID: PMC2238833

PMID: 17921498 [Indexed for MEDLINE]

2562. Nucleic Acids Res. 2008 Jan;36(Database issue):D303-6. Epub 2007 Oct 4.

DB-PABP: a database of polyanion-binding proteins.

Fang J(1), Dong Y, Salamat-Miller N, Middaugh CR.

Author information:

(1)Bioinformatics Core Facility, Information and Telecommunication Technology

Center, University of Kansas, Lawrence, KS 66047, USA.

The interactions between polyanions (PAs) and polyanion-binding proteins (PABPs)

have been found to play significant roles in many essential biological processes

including intracellular organization, transport and protein folding. Furthermore,

many neurodegenerative disease-related proteins are PABPs. Thus, a better

understanding of PA/PABP interactions may not only enhance our understandings of

biological systems but also provide new clues to these deadly diseases. The

literature in this field is widely scattered, suggesting the need for a

comprehensive and searchable database of PABPs. The DB-PABP is a comprehensive,

manually curated and searchable database of experimentally characterized PABPs.

It is freely available and can be accessed online at

http://pabp.bcf.ku.edu/DB\_PABP/. The DB-PABP was implemented as a MySQL

relational database. An interactive web interface was created using Java Server

Pages (JSP). The search page of the database is organized into a main search form

and a section for utilities. The main search form enables custom searches via

four menus: protein names, polyanion names, the source species of the proteins

and the methods used to discover the interactions. Available utilities include a

commonality matrix, a function of listing PABPs by the number of interacting

polyanions and a string search for author surnames. The DB-PABP is maintained at

the University of Kansas. We encourage users to provide feedback and submit new

data and references.

DOI: 10.1093/nar/gkm784

PMCID: PMC2238912

PMID: 17916573 [Indexed for MEDLINE]

2563. Pac Symp Biocomput. 2008:64-74.

Analysis of microRNA-target interactions by a target structure based

hybridization model.

Long D(1), Chan CY, Ding Y.

Author information:

(1)Wadsworth Center, New York State Department of Health, 150 New Scotland

Avenue, Albany, NY 12208, USA. dlong@wadsworth.org

MicroRNAs (miRNAs) are small non-coding RNAs that repress protein synthesis by

binding to target messenger RNAs (mRNAs) in multicellular eukaryotes. The

mechanism by which animal miRNAs specifically recognize their targets is not well

understood. We recently developed a model for modeling the interaction between a

miRNA and a target as a two-step hybridization reaction: nucleation at an

accessible target site, followed by hybrid elongation to disrupt local target

secondary structure and form the complete miRNA-target duplex. Nucleation

potential and hybridization energy are two key energetic characteristics of the

model. In this model, the role of target secondary structure on the efficacy of

repression by miRNAs is considered, by employing the Sfold program to address the

likelihood of a population of structures that co-exist in dynamic equilibrium for

a specific mRNA molecule. This model can accurately account for the sensitivity

to repression by let-7 of both published and rationally designed mutant forms of

the Caenorhabditis elegans lin-41 3' UTR, and for the behavior of many other

experimentally-tested miRNA-target interactions in C. elegans and Drosophila

melanogaster. The model is particularly effective in accounting for certain false

positive predictions obtained by other methods. In this study, we employed this

model to analyze a set of miRNA-target interactions that were experimentally

tested in mammalian models. These include targets for both mammalian miRNAs and

viral miRNAs, and a viral target of a human miRNA. We found that our model can

well account for both positive interactions and negative interactions. The model

provides a unique explanation for the lack of function of a conserved seed site

in the 3' UTR of the viral target, and predicts a strong interaction that cannot

be predicted by conservation-based methods. Thus, the findings from this analysis

and the previous analysis suggest that target structural accessibility is

generally important for miRNA function in a broad class of eukaryotic systems.

The model can be combined with other algorithms to improve the specificity of

predictions by these algorithms. Because the model does not involve sequence

conservation, it is readily applicable to target identification for microRNAs

that lack conserved sites, non-conserved human miRNAs, and poorly conserved viral

mRNAs. StarMir is a new Sfold application module developed for the implementation

of the structure-based model, and is available through Sfold Web server at

http://sfold.wadsworth.org.

PMID: 18232104 [Indexed for MEDLINE]

2564. PLoS Negl Trop Dis. 2008;2(10):e315. doi: 10.1371/journal.pntd.0000315. Epub 2008

Oct 22.

Web-based virtual microscopy for parasitology: a novel tool for education and

quality assurance.

Linder E(1), Lundin M, Thors C, Lebbad M, Winiecka-Krusnell J, Helin H, Leiva B,

Isola J, Lundin J.

Author information:

(1)Swedish Institute for Infectious Disease Control (SMI), Solna, Department of

Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, Stockholm,

Sweden. ewert.linder@smi.ki.se

Erratum in

PLoS Negl Trop Dis.

2008;2(10):10.1371/annotation/1f73ee39-9e3c-4ce4-9c35-2a6ab393de7d..

BACKGROUND: The basis for correctly assessing the burden of parasitic infections

and the effects of interventions relies on a somewhat shaky foundation as long as

we do not know how reliable the reported laboratory findings are. Thus virtual

microscopy, successfully introduced as a histopathology tool, has been adapted

for medical parasitology.

METHODOLOGY/PRINCIPAL FINDINGS: Specimens containing parasites in tissues,

stools, and blood have been digitized and made accessible as a "webmicroscope for

parasitology" (WMP) on the Internet

(http://www.webmicroscope.net/parasitology).These digitized specimens can be

viewed ("navigated" both in the x-axis and the y-axis) at the desired

magnification by an unrestricted number of individuals simultaneously. For

virtual microscopy of specimens containing stool parasites, it was necessary to

develop the technique further in order to enable navigation in the z plane (i.e.,

"focusing"). Specimens were therefore scanned and photographed in two or more

focal planes. The resulting digitized specimens consist of stacks of laterally

"stiched" individual images covering the entire area of the sample photographed

at high magnification. The digitized image information (approximately 10 GB

uncompressed data per specimen) is accessible at data transfer speeds from 2 to

10 Mb/s via a network of five image servers located in different parts of Europe.

Image streaming and rapid data transfer to an ordinary personal computer makes

web-based virtual microscopy similar to conventional microscopy.

CONCLUSION/SIGNIFICANCE: The potential of this novel technique in the field of

medical parasitology to share identical parasitological specimens means that we

can provide a "gold standard", which can overcome several problems encountered in

quality control of diagnostic parasitology. Thus, the WMP may have an impact on

the reliability of data, which constitute the basis for our understanding of the

vast problem of neglected tropical diseases. The WMP can be used also in the

absence of a fast Internet communication. An ordinary PC, or even a laptop, may

function as a local image server, e.g., in health centers in tropical endemic

areas.

DOI: 10.1371/journal.pntd.0000315

PMCID: PMC2565642

PMID: 18941514 [Indexed for MEDLINE]

2565. PLoS One. 2008;3(10):e3375. doi: 10.1371/journal.pone.0003375. Epub 2008 Oct 10.

Probing metagenomics by rapid cluster analysis of very large datasets.

Li W(1), Wooley JC, Godzik A.

Author information:

(1)California Institute for Telecommunications and Information Technology,

University of California San Diego, La Jolla, California, USA.

BACKGROUND: The scale and diversity of metagenomic sequencing projects challenge

both our technical and conceptual approaches in gene and genome annotations. The

recent Sorcerer II Global Ocean Sampling (GOS) expedition yielded millions of

predicted protein sequences, which significantly altered the landscape of known

protein space by more than doubling its size and adding thousands of new families

(Yooseph et al., 2007 PLoS Biol 5, e16). Such datasets, not only by their sheer

size, but also by many other features, defy conventional analysis and annotation

methods.

METHODOLOGY/PRINCIPAL FINDINGS: In this study, we describe an approach for rapid

analysis of the sequence diversity and the internal structure of such very large

datasets by advanced clustering strategies using the newly modified CD-HIT

algorithm. We performed a hierarchical clustering analysis on the 17.4 million

Open Reading Frames (ORFs) identified from the GOS study and found over 33

thousand large predicted protein clusters comprising nearly 6 million sequences.

Twenty percent of these clusters did not match known protein families by sequence

similarity search and might represent novel protein families. Distributions of

the large clusters were illustrated on organism composition, functional class,

and sample locations.

CONCLUSION/SIGNIFICANCE: Our clustering took about two orders of magnitude less

computational effort than the similar protein family analysis of original GOS

study. This approach will help to analyze other large metagenomic datasets in the

future. A Web server with our clustering results and annotations of predicted

protein clusters is available online at http://tools.camera.calit2.net/gos under

the CAMERA project.

DOI: 10.1371/journal.pone.0003375

PMCID: PMC2557142

PMID: 18846219 [Indexed for MEDLINE]

2566. Protein Pept Lett. 2008;15(9):956-63.

TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic

disorder.

Campen A(1), Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK.

Author information:

(1)Department of Computer and Information Science, Indiana University-Purdue

University Indianapolis, IN 46202, USA.

Intrinsically disordered proteins carry out various biological functions while

lacking ordered secondary and/or tertiary structure. In order to find general

intrinsic properties of amino acid residues that are responsible for the absence

of ordered structure in intrinsically disordered proteins we surveyed 517 amino

acid scales. Each of these scales was taken as an independent attribute for the

subsequent analysis. For a given attribute value X, which is averaged over a

consecutive string of amino acids, and for a given data set having both ordered

and disordered segments, the conditional probabilities P(s(o) | x) and P(s(d) |

x) for order and disorder, respectively, can be determined for all possible

values of X. Plots of the conditional probabilities P(s(o) | x) and P(s(o) | x)

versus X give a pair of curves. The area between these two curves divided by the

total area of the graph gives the area ratio value (ARV), which is proportional

to the degree of separation of the two probability curves and, therefore,

provides a measure of the given attribute's power to discriminate between order

and disorder. As ARV falls between zero and one, larger ARV corresponds to the

better discrimination between order and disorder. Starting from the scale with

the highest ARV, we applied a simulated annealing procedure to search for

alternative scale values and have managed to increase the ARV by more than 10%.

The ranking of the amino acids in this new TOP-IDP scale is as follows (from

order promoting to disorder promoting): W, F, Y, I, M, L, V, N, C, T, A, G, R, D,

H, Q, K, S, E, P. A web-based server has been created to apply the TOP-IDP scale

to predict intrinsically disordered proteins

(http://www.disprot.org/dev/disindex.php).

PMCID: PMC2676888

PMID: 18991772 [Indexed for MEDLINE]

2567. Protein Pept Lett. 2008;15(6):590-4.

Predicting membrane protein types with bragging learner.

Niu B(1), Jin YH, Feng KY, Liu L, Lu WC, Cai YD, Li GZ.

Author information:

(1)School of Materials Science and Engineering, Shanghai University, Shanghai,

People's Republic of China.

The membrane protein type is an important feature in characterizing the overall

topological folding type of a protein or its domains therein. Many investigators

have put their efforts to the prediction of membrane protein type. Here, we

propose a new approach, the bootstrap aggregating method or bragging learner, to

address this problem based on the protein amino acid composition. As a

demonstration, the benchmark dataset constructed by K.C. Chou and D.W. Elrod was

used to test the new method. The overall success rate thus obtained by jackknife

cross-validation was over 84%, indicating that the bragging learner as presented

in this paper holds a quite high potential in predicting the attributes of

proteins, or at least can play a complementary role to many existing algorithms

in this area. It is anticipated that the prediction quality can be further

enhanced if the pseudo amino acid composition can be effectively incorporated

into the current predictor. An online membrane protein type prediction web server

developed in our lab is available at

http://chemdata.shu.edu.cn/protein/protein.jsp.

PMID: 18680454 [Indexed for MEDLINE]

2568. Protein Pept Lett. 2008;15(3):286-9.

Predicting subcellular localization with AdaBoost Learner.

Jin YH(1), Niu B, Feng KY, Lu WC, Cai YD, Li GZ.

Author information:

(1)Department of Chemistry, College of Sciences, Shanghai University, 99 Shang-Da

Road, Shanghai, China 200444.

Protein subcellular localization, which tells where a protein resides in a cell,

is an important characteristic of a protein, and relates closely to the function

of proteins. The prediction of their subcellular localization plays an important

role in the prediction of protein function, genome annotation and drug design.

Therefore, it is an important and challenging role to predict subcellular

localization using bio-informatics approach. In this paper, a robust predictor,

AdaBoost Learner is introduced to predict protein subcellular localization based

on its amino acid composition. Jackknife cross-validation and independent dataset

test were used to demonstrate that Adaboost is a robust and efficient model in

predicting protein subcellular localization. As a result, the correct prediction

rates were 74.98% and 80.12% for the Jackknife test and independent dataset test

respectively, which are higher than using other existing predictors. An online

server for predicting subcellular localization of proteins based on AdaBoost

classifier was available on http://chemdata.shu. edu.cn/sl12.

PMID: 18336359 [Indexed for MEDLINE]

2569. Protein Pept Lett. 2008;15(1):33-8.

COILCHECK: an interactive server for the analysis of interface regions in coiled

coils.

Alva V(1), Syamala Devi DP, Sowdhamini R.

Author information:

(1)Institute of Bioinformatics and Applied Biotechnology, G-05, Tech Park Mall,

Internatioonal Technology Park Bangalore, Whitefield Road, Bangalore 560 066,

India.

Coiled coils are important structural modules and domains of protein-protein

interactions. Further, they provide the required framework to proteins, like

kinesins, myosins and SNAREs, which are either structural or involved in

transport of biomolecules. We provide an interactive webserver to measure the

strength of interactions between two helices involved in coiled coils.

Interactions are measured using non-bonded and electrostatic interactions and the

presence of hydrogen bonds and salt bridges. The sum of these interactions is

expressed as psuedoenergy, whose ranges have been standardized using all

structural entries that are classified to contain coiled coils. The results are

displayed conveniently as energy per residue along with options to obtain

detailed list of different types of interactions. This webserver can be useful to

assess the strength of coiled coil regions, to recognize weak and strong regions,

to rationalize the phenotypic behaviour of single residue mutations as well as to

design mutation experiments. COILCHECK webserver can be accessed from

http://caps.ncbs.res.in/coilcheck/.

PMID: 18221010 [Indexed for MEDLINE]

2570. SAR QSAR Environ Res. 2008 Jan-Mar;19(1-2):1-9. doi: 10.1080/10629360701843540.

Internet resources integrating many small-molecule databases.

Sitzmann M(1), Filippov IV, Nicklaus MC.

Author information:

(1)Computer-Aided Drug Design Group, Laboratory of Medicinal Chemistry, Center

for Cancer Research, National Cancer Institute, National Institutes of Health,

DHHS, Frederick, MD, USA.

()New data, tools and services recently made available on the web server

(http://cactus.nci.nih.gov) of the Computer-Aided Drug Design (CADD) Group, NCI,

NIH, developed in the context of chemoinformatics and drug development work, are

presented. These tools are designed for searching for structures in very large

databases of small molecules. One of them is a web service-the Chemical Structure

Look-up Service (CSLS)-for very rapid structure look-up in an aggregated

collection of more than 80 databases comprising more than 27 million unique

structures at the time of this writing. CSLS contains pointers to the entries in

toxicology-related databases, catalogues of commercially available samples,

drugs, assay results data sets, and databases in several other categories. CSLS

allows the user to find out very rapidly in which one(s) of all these databases a

given structure occurs independent of the representation of the input structure,

by making use of InChIs as well as new CACTVS hashcode-based identifiers. These

latter, calculable, identifiers are designed to take into account tautomerism,

different resonance structures drawn for charged species, and presence of

additional fragments. They make possible fine-tunable yet rapid compound

identification and database overlap analyses in very large compound collections.

DOI: 10.1080/10629360701843540

PMID: 18311630 [Indexed for MEDLINE]

2571. Bioinformatics. 2007 Dec 15;23(24):3397-9. Epub 2007 Oct 12.

InterProSurf: a web server for predicting interacting sites on protein surfaces.

Negi SS(1), Schein CH, Oezguen N, Power TD, Braun W.

Author information:

(1)Department of Biochemistry and Molecular Biology, Sealy Center for Structural

Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston,

Texas 77555-0857, USA.

A new web server, InterProSurf, predicts interacting amino acid residues in

proteins that are most likely to interact with other proteins, given the 3D

structures of subunits of a protein complex. The prediction method is based on

solvent accessible surface area of residues in the isolated subunits, a

propensity scale for interface residues and a clustering algorithm to identify

surface regions with residues of high interface propensities. Here we illustrate

the application of InterProSurf to determine which areas of Bacillus anthracis

toxins and measles virus hemagglutinin protein interact with their respective

cell surface receptors. The computationally predicted regions overlap with those

regions previously identified as interface regions by sequence analysis and

mutagenesis experiments.AVAILABILITY: The InterProSurf web server is available at

http://curie.utmb.edu/

DOI: 10.1093/bioinformatics/btm474

PMCID: PMC2636624

PMID: 17933856 [Indexed for MEDLINE]

2572. Bioinformatics. 2007 Dec 15;23(24):3403-5. Epub 2007 Oct 5.

XtalPred: a web server for prediction of protein crystallizability.

Slabinski L(1), Jaroszewski L, Rychlewski L, Wilson IA, Lesley SA, Godzik A.

Author information:

(1)Joint Center for Structural Genomics, La Jolla, CA 92037, USA.

XtalPred is a web server for prediction of protein crystallizability. The

prediction is made by comparing several features of the protein with

distributions of these features in TargetDB and combining the results into an

overall probability of crystallization. XtalPred provides: (1) a detailed

comparison of the protein's features to the corresponding distribution from

TargetDB; (2) a summary of protein features and predictions that indicate

problems that are likely to be encountered during protein crystallization; (3)

prediction of ligands; and (4) (optional) lists of close homologs from complete

microbial genomes that are more likely to crystallize.AVAILABILITY: The XtalPred

web server is freely available for academic users on

http://ffas.burnham.org/XtalPred

DOI: 10.1093/bioinformatics/btm477

PMID: 17921170 [Indexed for MEDLINE]

2573. Bioinformatics. 2007 Dec 15;23(24):3386-7. Epub 2007 Sep 25.

meta-PPISP: a meta web server for protein-protein interaction site prediction.

Qin S(1), Zhou HX.

Author information:

(1)Institute of Molecular Biophysics, School of Computational Science, Florida

State University, Tallahassee, Florida 32306, USA.

A number of complementary methods have been developed for predicting

protein-protein interaction sites. We sought to increase prediction robustness

and accuracy by combining results from different predictors, and report here a

meta web server, meta-PPISP, that is built on three individual web servers:

cons-PPISP (http://pipe.scs.fsu.edu/ppisp.html), Promate

(http://bioportal.weizmann.ac.il/promate), and PINUP

(http://sparks.informatics.iupui.edu/PINUP/). A linear regression method, using

the raw scores of the three servers as input, was trained on a set of 35

nonhomologous proteins. Cross validation showed that meta-PPISP outperforms all

the three individual servers. At coverages identical to those of the individual

methods, the accuracy of meta-PPISP is higher by 4.8 to 18.2 percentage points.

Similar improvements in accuracy are also seen on CAPRI and other

targets.AVAILABILITY: meta-PPISP can be accessed at

http://pipe.scs.fsu.edu/meta-ppisp.html

DOI: 10.1093/bioinformatics/btm434

PMID: 17895276 [Indexed for MEDLINE]

2574. BMC Bioinformatics. 2007 Dec 3;8:471.

Predicting and improving the protein sequence alignment quality by support vector

regression.

Lee M(1), Jeong CS, Kim D.

Author information:

(1)Department of Bio and Brain Engineering, Korea Advanced Institute of Science

and Technology, Daejeon, Republic of Korea. MinhoLee@kaist.edu

BACKGROUND: For successful protein structure prediction by comparative modeling,

in addition to identifying a good template protein with known structure,

obtaining an accurate sequence alignment between a query protein and a template

protein is critical. It has been known that the alignment accuracy can vary

significantly depending on our choice of various alignment parameters such as gap

opening penalty and gap extension penalty. Because the accuracy of sequence

alignment is typically measured by comparing it with its corresponding structure

alignment, there is no good way of evaluating alignment accuracy without knowing

the structure of a query protein, which is obviously not available at the time of

structure prediction. Moreover, there is no universal alignment parameter option

that would always yield the optimal alignment.

RESULTS: In this work, we develop a method to predict the quality of the

alignment between a query and a template. We train the support vector regression

(SVR) models to predict the MaxSub scores as a measure of alignment quality. The

alignment between a query protein and a template of length n is transformed into

a (n + 1)-dimensional feature vector, then it is used as an input to predict the

alignment quality by the trained SVR model. Performance of our work is evaluated

by various measures including Pearson correlation coefficient between the

observed and predicted MaxSub scores. Result shows high correlation coefficient

of 0.945. For a pair of query and template, 48 alignments are generated by

changing alignment options. Trained SVR models are then applied to predict the

MaxSub scores of those and to select the best alignment option which is chosen

specifically to the query-template pair. This adaptive selection procedure

results in 7.4% improvement of MaxSub scores, compared to those when the single

best parameter option is used for all query-template pairs.

CONCLUSION: The present work demonstrates that the alignment quality can be

predicted with reasonable accuracy. Our method is useful not only for selecting

the optimal alignment parameters for a chosen template based on predicted

alignment quality, but also for filtering out problematic templates that are not

suitable for structure prediction due to poor alignment accuracy. This is

implemented as a part in FORECAST, the server for fold-recognition and is freely

available on the web at http://pbil.kaist.ac.kr/forecast.

DOI: 10.1186/1471-2105-8-471

PMCID: PMC2222655

PMID: 18053160 [Indexed for MEDLINE]

2575. Bioinformatics. 2007 Dec 1;23(23):3244-6. Epub 2007 Oct 31.

Pepitope: epitope mapping from affinity-selected peptides.

Mayrose I(1), Penn O, Erez E, Rubinstein ND, Shlomi T, Freund NT, Bublil EM,

Ruppin E, Sharan R, Gershoni JM, Martz E, Pupko T.

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

Identifying the epitope to which an antibody binds is central for many

immunological applications such as drug design and vaccine development. The

Pepitope server is a web-based tool that aims at predicting discontinuous

epitopes based on a set of peptides that were affinity-selected against a

monoclonal antibody of interest. The server implements three different algorithms

for epitope mapping: PepSurf, Mapitope, and a combination of the two. The

rationale behind these algorithms is that the set of peptides mimics the genuine

epitope in terms of physicochemical properties and spatial organization. When the

three-dimensional (3D) structure of the antigen is known, the information in

these peptides can be used to computationally infer the corresponding epitope. A

user-friendly web interface and a graphical tool that allows viewing the

predicted epitopes were developed. Pepitope can also be applied for inferring

other types of protein-protein interactions beyond the immunological context, and

as a general tool for aligning linear sequences to a 3D structure.AVAILABILITY:

http://pepitope.tau.ac.il/

DOI: 10.1093/bioinformatics/btm493

PMID: 17977889 [Indexed for MEDLINE]

2576. Bioinformatics. 2007 Dec 1;23(23):3147-54. Epub 2007 Oct 17.

Predicting disulfide connectivity from protein sequence using multiple sequence

feature vectors and secondary structure.

Song J(1), Yuan Z, Tan H, Huber T, Burrage K.

Author information:

(1)Advanced Computational Modelling Centre, The University of Queensland,

Brisbane, QLD 4072, Australia.

MOTIVATION: Disulfide bonds are primary covalent crosslinks between two cysteine

residues in proteins that play critical roles in stabilizing the protein

structures and are commonly found in extracy-toplasmatic or secreted proteins. In

protein folding prediction, the localization of disulfide bonds can greatly

reduce the search in conformational space. Therefore, there is a great need to

develop computational methods capable of accurately predicting disulfide

connectivity patterns in proteins that could have potentially important

applications.

RESULTS: We have developed a novel method to predict disulfide connectivity

patterns from protein primary sequence, using a support vector regression (SVR)

approach based on multiple sequence feature vectors and predicted secondary

structure by the PSIPRED program. The results indicate that our method could

achieve a prediction accuracy of 74.4% and 77.9%, respectively, when averaged on

proteins with two to five disulfide bridges using 4-fold cross-validation,

measured on the protein and cysteine pair on a well-defined non-homologous

dataset. We assessed the effects of different sequence encoding schemes on the

prediction performance of disulfide connectivity. It has been shown that the

sequence encoding scheme based on multiple sequence feature vectors coupled with

predicted secondary structure can significantly improve the prediction accuracy,

thus enabling our method to outperform most of other currently available

predictors. Our work provides a complementary approach to the current algorithms

that should be useful in computationally assigning disulfide connectivity

patterns and helps in the annotation of protein sequences generated by

large-scale whole-genome projects.

AVAILABILITY: The prediction web server and Supplementary Material are accessible

at http://foo.maths.uq.edu.au/~huber/disulfide

DOI: 10.1093/bioinformatics/btm505

PMID: 17942444 [Indexed for MEDLINE]

2577. Bioinformatics. 2007 Dec 1;23(23):3241-3. Epub 2007 Jun 28.

CASVM: web server for SVM-based prediction of caspase substrates cleavage sites.

Wee LJ(1), Tan TW, Ranganathan S.

Author information:

(1)Department of Biochemistry, Yong Loo Lin School of Medicine, National

University of Singapore, Singapore.

Caspases belong to a unique class of cysteine proteases which function as

critical effectors of apoptosis, inflammation and other important cellular

processes. Caspases cleave substrates at specific tetrapeptide sites after a

highly conserved aspartic acid residue. Prediction of such cleavage sites will

complement structural and functional studies on substrates cleavage as well as

discovery of new substrates. We have recently developed a support vector machines

(SVM) method to address this issue. Our algorithm achieved an accuracy ranging

from 81.25 to 97.92%, making it one of the best methods currently available.

CASVM is the web server implementation of our SVM algorithms, written in Perl and

hosted on a Linux platform. The server can be used for predicting non-canonical

caspase substrate cleavage sites. We have also included a relational database

containing experimentally verified caspase substrates retrievable using accession

IDs, keywords or sequence similarity.AVAILABILITY:

http://www.casbase.org/casvm/index.html

DOI: 10.1093/bioinformatics/btm334

PMID: 17599937 [Indexed for MEDLINE]

2578. Genome Res. 2007 Dec;17(12):1797-808. Epub 2007 Nov 5.

28-way vertebrate alignment and conservation track in the UCSC Genome Browser.

Miller W(1), Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King

DC, Baertsch R, Blankenberg D, Kosakovsky Pond SL, Nekrutenko A, Giardine B,

Harris RS, Tyekucheva S, Diekhans M, Pringle TH, Murphy WJ, Lesk A, Weinstock GM,

Lindblad-Toh K, Gibbs RA, Lander ES, Siepel A, Haussler D, Kent WJ.

Author information:

(1)Center for Comparative Genomics and Bioinformatics, Penn State University,

University Park, Pennsylvania 16802, USA. webb@bx.psu.edu

This article describes a set of alignments of 28 vertebrate genome sequences that

is provided by the UCSC Genome Browser. The alignments can be viewed on the Human

Genome Browser (March 2006 assembly) at http://genome.ucsc.edu, downloaded in

bulk by anonymous FTP from

http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz28way, or analyzed with the

Galaxy server at http://g2.bx.psu.edu. This article illustrates the power of this

resource for exploring vertebrate and mammalian evolution, using three examples.

First, we present several vignettes involving insertions and deletions within

protein-coding regions, including a look at some human-specific indels. Then we

study the extent to which start codons and stop codons in the human sequence are

conserved in other species, showing that start codons are in general more poorly

conserved than stop codons. Finally, an investigation of the phylogenetic depth

of conservation for several classes of functional elements in the human genome

reveals striking differences in the rates and modes of decay in alignability.

Each functional class has a distinctive period of stringent constraint, followed

by decays that allow (for the case of regulatory regions) or reject (for coding

regions and ultraconserved elements) insertions and deletions.

DOI: 10.1101/gr.6761107

PMCID: PMC2099589

PMID: 17984227 [Indexed for MEDLINE]

2579. Genomics Proteomics Bioinformatics. 2007 Dec;5(3-4):250-2. doi:

10.1016/S1672-0229(08)60012-1.

Oxypred: prediction and classification of oxygen-binding proteins.

Muthukrishnan S(1), Garg A, Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector 39-A, Chandigarh 160036, India.

This study describes a method for predicting and classifying oxygen-binding

proteins. Firstly, support vector machine (SVM) modules were developed using

amino acid composition and dipeptide composition for predicting oxygen-binding

proteins, and achieved maximum accuracy of 85.5% and 87.8%, respectively.

Secondly, an SVM module was developed based on amino acid composition,

classifying the predicted oxygen-binding proteins into six classes with accuracy

of 95.8%, 97.5%, 97.5%, 96.9%, 99.4%, and 96.0% for erythrocruorin, hemerythrin,

hemocyanin, hemoglobin, leghemoglobin, and myoglobin proteins, respectively.

Finally, an SVM module was developed using dipeptide composition for classifying

the oxygen-binding proteins, and achieved maximum accuracy of 96.1%, 98.7%,

98.7%, 85.6%, 99.6%, and 93.3% for the above six classes, respectively. All

modules were trained and tested by five-fold cross validation. Based on the above

approach, a web server Oxypred was developed for predicting and classifying

oxygen-binding proteins (available from

http://www.imtech.res.in/raghava/oxypred/).

DOI: 10.1016/S1672-0229(08)60012-1

PMCID: PMC5054225

PMID: 18267306 [Indexed for MEDLINE]

2580. J Digit Imaging. 2007 Dec;20(4):393-401.

Online availability check of teleradiology components.

Weisser G(1), Ruggiero S, Runa A, Düber C, Neff W, Walz M.

Author information:

(1)Department of Clinical Radiology, University Hospital Mannheim,

Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany.

gerald.weisser@rad.ma.uni-heidelberg.de

For a region-wide teleradiology network in Germany a vendor-independent

Uptime-server concept was defined. The Uptime-server was realized for the

availability check and prospective error-detection of the emergency teleradiology

servers and clients based on encrypted digital imaging and communication in

medicine (DICOM)-e-mail transfers. The concept and the experiences of 2 years of

use with more than 30 clients and servers in 15 hospitals and in nine other

regional partners are shown. The Uptime-server does provide automated

availability checks for all servers and clients, automated checks of the download

speed of the Internet lines, and a graphical user interface for the clinical user

and the system administrator. A clinical user can display the availability

information from all clients and servers in the network (see

http://www.teleradiologie-rnd.de). In case of malfunctions during an emergency

transfer, immediate reactions are possible, often without the need for help of a

hotline or a system administrator. The chosen Uptime-server concept proofed to be

reliable; it worked with products from nine different manufacturers without

problems. Its statistical output can be used to fulfill the legal requirements of

regular availability checks for teleradiology lines.

DOI: 10.1007/s10278-006-1044-3

PMCID: PMC3043915

PMID: 17252170 [Indexed for MEDLINE]

2581. Structure. 2007 Dec;15(12):1567-76.

Modeling backbone flexibility improves protein stability estimation.

Yin S(1), Ding F, Dokholyan NV.

Author information:

(1)Department of Biochemistry and Biophysics, School of Medicine, University of

North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

In designing mutagenesis experiments, it is often crucial to know how certain

mutations will affect the structure and thermodynamic stability of the protein.

Here, we present a methodology, Eris, to efficiently and accurately compute the

stability changes of proteins upon mutations using our protein-modeling suite,

Medusa. We evaluate the stability changes upon mutations for 595 mutants from

five structurally unrelated proteins, and find significant correlations between

the predicted and experimental results. For cases when the high-resolution

protein structure is not available, we find that better predictions are obtained

by backbone structure prerelaxation. The advantage of our approach is that it is

based on physical descriptions of atomic interactions, and does not rely on

parameter training with available experimental protein stability data. Unlike

other methods, Eris also models the backbone flexibility, thereby allowing for

determination of the mutation-induced backbone conformational changes. Eris is

freely available via the web server at http://eris.dokhlab.org.

DOI: 10.1016/j.str.2007.09.024

PMID: 18073107 [Indexed for MEDLINE]

2582. BMC Bioinformatics. 2007 Nov 30;8:469.

A structural analysis of in vitro catalytic activities of hammerhead ribozymes.

Shao Y(1), Wu S, Chan CY, Klapper JR, Schneider E, Ding Y.

Author information:

(1)Wadsworth Center, New York State Department of Health, 150 New Scotland

Avenue, Albany, NY 12208, USA. yushao04@yahoo.com

BACKGROUND: Ribozymes are small catalytic RNAs that possess the dual functions of

sequence-specific RNA recognition and site-specific cleavage. Trans-cleaving

ribozymes can inhibit translation of genes at the messenger RNA (mRNA) level in

both eukaryotic and prokaryotic systems and are thus useful tools for studies of

gene function. However, identification of target sites for efficient cleavage

poses a challenge. Here, we have considered a number of structural and

thermodynamic parameters that can affect the efficiency of target cleavage, in an

attempt to identify rules for the selection of functional ribozymes.

RESULTS: We employed the Sfold program for RNA secondary structure prediction, to

account for the likely population of target structures that co-exist in dynamic

equilibrium for a specific mRNA molecule. We designed and prepared 15 hammerhead

ribozymes to target GUC cleavage sites in the mRNA of the breast cancer

resistance protein (BCRP). These ribozymes were tested, and their catalytic

activities were measured in vitro. We found that target disruption energy owing

to the alteration of the local target structure necessary for ribozyme binding,

and the total energy change of the ribozyme-target hybridization, are two

significant parameters for prediction of ribozyme activity. Importantly, target

disruption energy is the major contributor to the predictability of ribozyme

activity by the total energy change. Furthermore, for a target-site specific

ribozyme, incorrect folding of the catalytic core, or interactions involving the

two binding arms and the end sequences of the catalytic core, can have

detrimental effects on ribozyme activity.

CONCLUSION: The findings from this study suggest rules for structure-based

rational design of trans-cleaving hammerhead ribozymes in gene knockdown studies.

Tools implementing these rules are available from the Sribo module and the Srna

module of the Sfold program available through Web server at

http://sfold.wadsworth.org.

DOI: 10.1186/1471-2105-8-469

PMCID: PMC2238771

PMID: 18053134 [Indexed for MEDLINE]

2583. BMC Bioinformatics. 2007 Nov 27;8:463.

Identification of DNA-binding proteins using support vector machines and

evolutionary profiles.

Kumar M(1), Gromiha MM, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39A,

Chandigarh-160036, India. manish@imtech.res.in

BACKGROUND: Identification of DNA-binding proteins is one of the major challenges

in the field of genome annotation, as these proteins play a crucial role in

gene-regulation. In this paper, we developed various SVM modules for predicting

DNA-binding domains and proteins. All models were trained and tested on multiple

datasets of non-redundant proteins.

RESULTS: SVM models have been developed on DNAaset, which consists of 1153

DNA-binding and equal number of non DNA-binding proteins, and achieved the

maximum accuracy of 72.42% and 71.59% using amino acid and dipeptide

compositions, respectively. The performance of SVM model improved from 72.42% to

74.22%, when evolutionary information in form of PSSM profiles was used as input

instead of amino acid composition. In addition, SVM models have been developed on

DNAset, which consists of 146 DNA-binding and 250 non-binding chains/domains, and

achieved the maximum accuracy of 79.80% and 86.62% using amino acid composition

and PSSM profiles. The SVM models developed in this study perform better than

existing methods on a blind dataset.

CONCLUSION: A highly accurate method has been developed for predicting

DNA-binding proteins using SVM and PSSM profiles. This is the first study in

which evolutionary information in form of PSSM profiles has been used

successfully for predicting DNA-binding proteins. A web-server DNAbinder has been

developed for identifying DNA-binding proteins and domains from query amino acid

sequences http://www.imtech.res.in/raghava/dnabinder/.

DOI: 10.1186/1471-2105-8-463

PMCID: PMC2216048

PMID: 18042272 [Indexed for MEDLINE]

2584. Biochem Biophys Res Commun. 2007 Nov 16;363(2):297-303. Epub 2007 Aug 31.

Signal-3L: A 3-layer approach for predicting signal peptides.

Shen HB(1), Chou KC.

Author information:

(1)Gordon Life Science Institute, San Diego, CA 92130, USA.

hbshen@crystal.harvard.edu

Functioning as an "address tag" that directs nascent proteins to their proper

cellular and extracellular locations, signal peptides have become a crucial tool

in finding new drugs or reprogramming cells for gene therapy. To effectively and

timely use such a tool, however, the first important thing is to develop an

automated method for rapidly and accurately identifying the signal peptide for a

given nascent protein. With the avalanche of new protein sequences generated in

the post-genomic era, the challenge has become even more urgent and critical. In

this paper, we have developed a novel method for predicting signal peptide

sequences and their cleavage sites in human, plant, animal, eukaryotic,

Gram-positive, and Gram-negative protein sequences, respectively. The new

predictor is called Signal-3L that consists of three prediction engines working,

respectively, for the following three progressively deepening layers: (1)

identifying a query protein as secretory or non-secretory by an ensemble

classifier formed by fusing many individual OET-KNN (optimized evidence-theoretic

K nearest neighbor) classifiers operated in various dimensions of PseAA (pseudo

amino acid) composition spaces; (2) selecting a set of candidates for the

possible signal peptide cleavage sites of a query secretory protein by a

subsite-coupled discrimination algorithm; (3) determining the final cleavage site

by fusing the global sequence alignment outcome for each of the aforementioned

candidates through a voting system. Signal-3L is featured by high success

prediction rates with short computational time, and hence is particularly useful

for the analysis of large-scale datasets. Signal-3L is freely available as a

web-server at http://chou.med.harvard.edu/bioinf/Signal-3L/ or

http://202.120.37.186/bioinf/Signal-3L, where, to further support the demand of

the related areas, the signal peptides identified by Signal-3L for all the

protein entries in Swiss-Prot databank that do not have signal peptide

annotations or are annotated with uncertain terms but are classified by Signal-3L

as secretory proteins are provided in a downloadable file. The large-scale file

is prepared with Microsoft Excel and named "Tab-Signal-3L.xls", and will be

updated once a year to include new protein entries and reflect the continuous

development of Signal-3L.

DOI: 10.1016/j.bbrc.2007.08.140

PMID: 17880924 [Indexed for MEDLINE]

2585. J Photochem Photobiol B. 2007 Nov 12;89(1):29-35. Epub 2007 Aug 7.

Database on mycosporines and mycosporine-like amino acids (MAAs) in fungi,

cyanobacteria, macroalgae, phytoplankton and animals.

Sinha RP(1), Singh SP, Häder DP.

Author information:

(1)Centre of Advanced Study in Botany, Banaras Hindu University, Varanasi-221

005, India.

A database on UV-absorbing mycosporines and mycosporine-like amino acids (MAAs)

has been constructed that provides information on various mycosporines and MAAs

reported in fungi, cyanobacteria, macroalgae, phytoplankton and animals from

aquatic and terrestrial habitats. It also contains information on biosynthetic

routes of MAAs as well as on the absorption maxima and molecular structures of

different mycosporines and MAAs (Table 1S). This database provides necessary

information for scientists working in the field of photoprotective compounds in

fungi, cyanobacteria, macroalgae, phytoplankton and animals (Table 2S). (Tables

1S and 2S are available online as Supplementary material in the electronic copy

of the journal as well as on our server

<http://www.biologie.uni-erlangen.de/botanik1/html/eng/maa\_database.htm>.).

DOI: 10.1016/j.jphotobiol.2007.07.006

PMID: 17826148 [Indexed for MEDLINE]

2586. BMC Bioinformatics. 2007 Nov 6;8:429.

CoryneRegNet 4.0 - A reference database for corynebacterial gene regulatory

networks.

Baumbach J(1).

Author information:

(1)Computational Methods for Emerging Technologies, Bielefeld University,

Bielefeld, Germany. Jan.Baumbach@CeBiTec.Uni-Bielefeld.DE

BACKGROUND: Detailed information on DNA-binding transcription factors (the key

players in the regulation of gene expression) and on transcriptional regulatory

interactions of microorganisms deduced from literature-derived knowledge,

computer predictions and global DNA microarray hybridization experiments, has

opened the way for the genome-wide analysis of transcriptional regulatory

networks. The large-scale reconstruction of these networks allows the in silico

analysis of cell behavior in response to changing environmental conditions. We

previously published CoryneRegNet, an ontology-based data warehouse of

corynebacterial transcription factors and regulatory networks. Initially, it was

designed to provide methods for the analysis and visualization of the gene

regulatory network of Corynebacterium glutamicum.

RESULTS: Now we introduce CoryneRegNet release 4.0, which integrates data on the

gene regulatory networks of 4 corynebacteria, 2 mycobacteria and the model

organism Escherichia coli K12. As the previous versions, CoryneRegNet provides a

web-based user interface to access the database content, to allow various

queries, and to support the reconstruction, analysis and visualization of

regulatory networks at different hierarchical levels. In this article, we present

the further improved database content of CoryneRegNet along with novel analysis

features. The network visualization feature GraphVis now allows the inter-species

comparisons of reconstructed gene regulatory networks and the projection of gene

expression levels onto that networks. Therefore, we added stimulon data directly

into the database, but also provide Web Service access to the DNA microarray

analysis platform EMMA. Additionally, CoryneRegNet now provides a SOAP based Web

Service server, which can easily be consumed by other bioinformatics software

systems. Stimulons (imported from the database, or uploaded by the user) can be

analyzed in the context of known transcriptional regulatory networks to predict

putative contradictions or further gene regulatory interactions. Furthermore, it

integrates protein clusters by means of heuristically solving the weighted graph

cluster editing problem. In addition, it provides Web Service based access to up

to date gene annotation data from GenDB.

CONCLUSION: The release 4.0 of CoryneRegNet is a comprehensive system for the

integrated analysis of procaryotic gene regulatory networks. It is a versatile

systems biology platform to support the efficient and large-scale analysis of

transcriptional regulation of gene expression in microorganisms. It is publicly

available at http://www.CoryneRegNet.DE.

DOI: 10.1186/1471-2105-8-429

PMCID: PMC2194740

PMID: 17986320 [Indexed for MEDLINE]

2587. Bioinformatics. 2007 Nov 1;23(21):2823-8. Epub 2007 Oct 5.

OSCAR: one-class SVM for accurate recognition of cis-elements.

Jiang B(1), Zhang MQ, Zhang X.

Author information:

(1)MOE Key Laboratory of Bioinformatics, Bioinformatics Division,

TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China.

MOTIVATION: Traditional methods to identify potential binding sites of known

transcription factors still suffer from large number of false predictions. They

mostly use sequence information in a position-specific manner and neglect other

types of information hidden in the proximal promoter regions. Recent biological

and computational researches, however, suggest that there exist not only

locational preferences of binding, but also correlations between transcription

factors.

RESULTS: In this article, we propose a novel approach, OSCAR, which utilizes

one-class SVM algorithms, and incorporates multiple factors to aid the

recognition of transcription factor binding sites. Using both synthetic and real

data, we find that our method outperforms existing algorithms, especially in the

high sensitivity region. The performance of our method can be further improved by

taking into account locational preference of binding events. By testing on

experimentally-verified binding sites of GATA and HNF transcription factor

families, we show that our algorithm can infer the true co-occurring motif pairs

accurately, and by considering the co-occurrences of correlated motifs, we not

only filter out false predictions, but also increase the sensitivity.

AVAILABILITY: An online server based on OSCAR is available at

http://bioinfo.au.tsinghua.edu.cn/oscar.

DOI: 10.1093/bioinformatics/btm473

PMID: 17921174 [Indexed for MEDLINE]

2588. Bioinformatics. 2007 Nov 1;23(21):2829-35. Epub 2007 Sep 25.

Minimizing the overlap problem in protein NMR: a computational framework for

precision amino acid labeling.

Sweredoski MJ(1), Donovan KJ, Nguyen BD, Shaka AJ, Baldi P.

Author information:

(1)Department of Computer Science, Institute for Genomics and Bioinformatics,

University of California, Irvine, USA.

MOTIVATION: Recent advances in cell-free protein expression systems allow

specific labeling of proteins with amino acids containing stable isotopes ((15)N,

(13) C and (2)H), an important feature for protein structure determination by

nuclear magnetic resonance (NMR) spectroscopy. Given this labeling ability, we

present a mathematical optimization framework for designing a set of protein

isotopomers, or labeling schedules, to reduce the congestion in the NMR spectra.

The labeling schedules, which are derived by the optimization of a cost function,

are tailored to a specific protein and NMR experiment.

RESULTS: For 2D (15)N-(1)H HSQC experiments, we can produce an exact solution

using a dynamic programming algorithm in under 2 h on a standard desktop machine.

Applying the method to a standard benchmark protein, calmodulin, we are able to

reduce the number of overlaps in the 500 MHz HSQC spectrum from 10 to 1 using

four samples with a true cost function, and 10 to 4 if the cost function is

derived from statistical estimates. On a set of 448 curated proteins from the

BMRB database, we are able to reduce the relative percent congestion by 84.9% in

their HSQC spectra using only four samples. Our method can be applied in a

high-throughput manner on a proteomic scale using the server we developed. On a

100-node cluster, optimal schedules can be computed for every protein coded for

in the human genome in less than a month.

AVAILABILITY: A server for creating labeling schedules for (15)N-(1)H HSQC

experiments as well as results for each of the individual 448 proteins used in

the test set is available at http://nmr.proteomics.ics.uci.edu.

DOI: 10.1093/bioinformatics/btm406

PMID: 17895278 [Indexed for MEDLINE]

2589. Bioinformatics. 2007 Nov 1;23(21):2945-6. Epub 2007 Sep 24.

Idiographica: a general-purpose web application to build idiograms on-demand for

human, mouse and rat.

Kin T(1), Ono Y.

Author information:

(1)Computational Biology Research Center (CBRC), National Institute of Advanced

Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064,

Japan. kin-taishin@aist.go.jp

SUMMARY: We have launched a web server, which serves as a general-purpose

idiogram rendering service, and allows users to generate high-quality idiograms

with custom annotation according to their own genome-wide mapping/annotation data

through an easy-to-use interface. The generated idiograms are suitable not only

for visualizing summaries of genome-wide analysis but also for many types of

presentation material including web pages, conference posters, oral

presentations, etc.

AVAILABILITY: Idiographica is freely available at

http://www.ncrna.org/idiographica/

DOI: 10.1093/bioinformatics/btm455

PMID: 17893084 [Indexed for MEDLINE]

2590. Bioinformatics. 2007 Nov 1;23(21):2959-60. Epub 2007 Sep 13.

RagPools: RNA-As-Graph-Pools--a web server for assisting the design of structured

RNA pools for in vitro selection.

Kim N(1), Shin JS, Elmetwaly S, Gan HH, Schlick T.

Author information:

(1)Department of Chemistry, New York University, 100 Washington Square East, New

York 10003, USA.

SUMMARY: Our RNA-As-Graph-Pools (RagPools) web server offers a theoretical

companion tool for RNA in vitro selection and related problems. Specifically, it

suggests how to construct RNA sequence/structure pools with user-specified

properties and assists in analyzing resulting distributions. This utility follows

our recently developed approach for engineering sequence pools that links RNA

sequence space regions with corresponding structural distributions via a 'mixing

matrix' approach combined with a graph theory analysis of RNA secondary-structure

space; the mixing matrix specifies nucleotide transition rates, and graph theory

links sequences to simple graphical objects representing RNA motifs. The

companion RagPools web server ('Designer' component) provides optimized starting

sequences, mixing matrices and associated weights in response to a user-specified

target pool structure distribution. In addition, RagPools ('Analyzer' component)

analyzes the motif distribution of pools generated from user-specified starting

sequences and mixing matrices. Thus, RagPools serves as a guide to researchers

who aim to synthesize RNA pools with desired properties and/or experiment in

silico with various designs by our approach.

AVAILABILITY: The web server is accessible on the web at

http://rubin2.biomath.nyu.edu

DOI: 10.1093/bioinformatics/btm439

PMID: 17855416 [Indexed for MEDLINE]

2591. Bioinformatics. 2007 Nov 1;23(21):2947-8. Epub 2007 Sep 10.

Clustal W and Clustal X version 2.0.

Larkin MA(1), Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H,

Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG.

Author information:

(1)The Conway Institute of Biomolecular and Biomedical Research, University

College Dublin, Belfield, Dublin 4, Ireland.

SUMMARY: The Clustal W and Clustal X multiple sequence alignment programs have

been completely rewritten in C++. This will facilitate the further development of

the alignment algorithms in the future and has allowed proper porting of the

programs to the latest versions of Linux, Macintosh and Windows operating

systems.

AVAILABILITY: The programs can be run on-line from the EBI web server:

http://www.ebi.ac.uk/tools/clustalw2. The source code and executables for

Windows, Linux and Macintosh computers are available from the EBI ftp site

ftp://ftp.ebi.ac.uk/pub/software/clustalw2/

DOI: 10.1093/bioinformatics/btm404

PMID: 17846036 [Indexed for MEDLINE]

2592. Bioinformation. 2007 Nov 1;2(3):86-90.

IWS: integrated web server for protein sequence and structure analysis.

Shameer K(1), Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences (TIFR), GKVK Campus, Bellary Road,

Bangalore 560 065, India.

Rapid increase in protein sequence information from genome sequencing projects

demand the intervention of bioinformatics tools to recognize interesting

gene-products and associated function. Often, multiple algorithms need to be

employed to improve accuracy in predictions and several structure prediction

algorithms are on the public domain. Here, we report the availability of an

Integrated Web-server as a bioinformatics online package dedicated for in-silico

analysis of protein sequence and structure data (IWS). IWS provides web interface

to both in-house and widely accepted programs from major bioinformatics groups,

organized as 10 different modules. IWS also provides interactive images for

Analysis Work Flow, which will provide transparency to the user to carry out

analysis by moving across modules seamlessly and to perform their predictions in

a rapid manner.AVAILABILITY: IWS IS AVAILABLE FROM THE URL:

http://caps.ncbs.res.in/iws.

PMCID: PMC2248443

PMID: 18288329

2593. Comput Biol Med. 2007 Nov;37(11):1672-5. Epub 2007 May 25.

A web server for automatic analysis and extraction of relevant biological

knowledge.

Cedano J(1), Huerta M, Estrada I, Ballllosera F, Conchillo O, Delicado P, Querol

E.

Author information:

(1)Institut de Biotecnologia i Biomedicina and Departament de Bioquímica i

Biología Molecular, UAB, Spain. jcedano@servet.uab.es

MOTIVATION: This application aims at assisting researchers with the extraction of

significant medical and biological knowledge from data sets with complex

relationships among their variables.

RESULTS: Non-hypothesis-driven approaches like Principal Curves of Oriented

Points (PCOP) are a very suitable method for this objective. PCOP allows for

obtaining of a representative pattern from a huge quantity of data of independent

variables in a very flexible and direct way. A web server has been designed to

automatically realize 'non-linear pattern' analysis, 'hidden-variable-dependent'

clustering, and new samples 'local-dispersion-dependent' classification from the

data involving new statistical techniques using the PCOP calculus. The tools

facilitate the managing, comparison and visualization of results in a

user-friendly graphical interface.

AVAILABILITY: http://ibb.uab.es/revresearch.

DOI: 10.1016/j.compbiomed.2007.03.008

PMID: 17531966 [Indexed for MEDLINE]

2594. J Mol Model. 2007 Nov;13(11):1157-67. Epub 2007 Sep 9.

Statistical analysis of physical-chemical properties and prediction of

protein-protein interfaces.

Negi SS(1), Braun W.

Author information:

(1)Department of Biochemistry and Molecular Biology, Sealy Center for Structural

Biology and Molecular Biophysics, University of Texas Medical Branch, 301

University Blvd, Galveston, TX 77555-0857, USA.

We have developed a fully automated method, InterProSurf, to predict interacting

amino acid residues on protein surfaces of monomeric 3D structures. Potential

interacting residues are predicted based on solvent accessible surface areas, a

new scale for interface propensities, and a cluster algorithm to locate surface

exposed areas with high interface propensities. Previous studies have shown the

importance of hydrophobic residues and specific charge distribution as

characteristics for interfaces. Here we show differences in interface and surface

regions of all physical chemical properties of residues as represented by five

quantitative descriptors. In the current study a set of 72 protein complexes with

known 3D structures were analyzed to obtain interface propensities of residues,

and to find differences in the distribution of five quantitative descriptors for

amino acid residues. We also investigated spatial pair correlations of solvent

accessible residues in interface and surface areas, and compared log-odds ratios

for interface and surface areas. A new scoring method to predict potential

functional sites on the protein surface was developed and tested for a new

dataset of 21 protein complexes, which were not included in the original training

dataset. Empirically we found that the algorithm achieves a good balance in the

accuracy of precision and sensitivity by selecting the top eight highest scoring

clusters as interface regions. The performance of the method is illustrated for a

dimeric ATPase of the hyperthermophile, Methanococcus jannaschii, and the capsid

protein of Human Hepatitis B virus. An automated version of the method can be

accessed from our web server at http://curie.utmb.edu/prosurf.html.

DOI: 10.1007/s00894-007-0237-0

PMCID: PMC2628805

PMID: 17828612 [Indexed for MEDLINE]

2595. Protein Eng Des Sel. 2007 Nov;20(11):561-7. Epub 2007 Nov 10.

Nuc-PLoc: a new web-server for predicting protein subnuclear localization by

fusing PseAA composition and PsePSSM.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, 1954 Hua-Shan Road, Shanghai 200030, China.

hbshen@crystal.harvard.edu

The life processes of an eukaryotic cell are guided by its nucleus. In addition

to the genetic material, the cellular nucleus contains many proteins located at

its different compartments, called subnuclear locations. Information of their

localization in a nucleus is indispensable for the in-depth study of system

biology because, in addition to helping determine their functions, it can provide

illuminative insights of how and in what kind of microenvironments these

subnuclear proteins are interacting with each other and with other molecules.

Facing the deluge of protein sequences generated in the post-genomic age, we are

challenged to develop an automated method for fast and effectively annotating the

subnuclear locations of numerous newly found nuclear protein sequences. In view

of this, a new classifier, called Nuc-PLoc, has been developed that can be used

to identify nuclear proteins among the following nine subnuclear locations: (1)

chromatin, (2) heterochromatin, (3) nuclear envelope, (4) nuclear matrix, (5)

nuclear pore complex, (6) nuclear speckle, (7) nucleolus, (8) nucleoplasm and (9)

nuclear promyelocytic leukaemia (PML) body. Nuc-PLoc is featured by an ensemble

classifier formed by fusing the evolution information of a protein and its

pseudo-amino acid composition. The overall jackknife cross-validation accuracy

obtained by Nuc-PLoc is significantly higher than those by the existing methods

on the same benchmark data set through the same testing procedure. As a

user-friendly web-server, Nuc-PLoc is freely accessible to the public at

http://chou.med.harvard.edu/bioinf/Nuc-PLoc.

DOI: 10.1093/protein/gzm057

PMID: 17993650 [Indexed for MEDLINE]

2596. BMC Bioinformatics. 2007 Oct 31;8:425.

SABERTOOTH: protein structural alignment based on a vectorial structure

representation.

Teichert F(1), Bastolla U, Porto M.

Author information:

(1)Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstr,

6-8, 64289 Darmstadt, Germany. floteich@fkp.tu-darmstadt.de

BACKGROUND: The task of computing highly accurate structural alignments of

proteins in very short computation time is still challenging. This is partly due

to the complexity of protein structures. Therefore, instead of manipulating

coordinates directly, matrices of inter-atomic distances, sets of vectors between

protein backbone atoms, and other reduced representations are used. These

decrease the effort of comparing large sets of coordinates, but protein

structural alignment still remains computationally expensive.

RESULTS: We represent the topology of a protein structure through a structural

profile that expresses the global effective connectivity of each residue. We have

shown recently that this representation allows explicitly expressing the

relationship between protein structure and protein sequence. Based on this very

condensed vectorial representation, we develop a structural alignment framework

that recognizes structural similarities with accuracy comparable to established

alignment tools. Furthermore, our algorithm has favourable scaling of computation

time with chain length. Since the algorithm is independent of the details of the

structural representation, our framework can be applied to sequence-to-sequence

and sequence-to-structure comparison within the same setup, and it is therefore

more general than other existing tools.

CONCLUSION: We show that protein comparison based on a vectorial representation

of protein structure performs comparably to established algorithms based on

coordinates. The conceptually new approach presented in this publication might

assist to unify the view on protein comparison by unifying structure and sequence

descriptions in this context. The framework discussed here is implemented in the

'SABERTOOTH' alignment server, freely accessible at

http://www.fkp.tu-darmstadt.de/sabertooth/.

DOI: 10.1186/1471-2105-8-425

PMCID: PMC2257979

PMID: 17974011 [Indexed for MEDLINE]

2597. BMC Bioinformatics. 2007 Oct 22;8:404.

Application of amino acid occurrence for discriminating different folding types

of globular proteins.

Taguchi YH(1), Gromiha MM.

Author information:

(1)Department of Physics, Faculty of Science and Technology, Chuo University,

1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. tag@granular.com

BACKGROUND: Predicting the three-dimensional structure of a protein from its

amino acid sequence is a long-standing goal in computational/molecular biology.

The discrimination of different structural classes and folding types are

intermediate steps in protein structure prediction.

RESULTS: In this work, we have proposed a method based on linear discriminant

analysis (LDA) for discriminating 30 different folding types of globular proteins

using amino acid occurrence. Our method was tested with a non-redundant set of

1612 proteins and it discriminated them with the accuracy of 38%, which is

comparable to or better than other methods in the literature. A web server has

been developed for discriminating the folding type of a query protein from its

amino acid sequence and it is available at http://granular.com/PROLDA/.

CONCLUSION: Amino acid occurrence has been successfully used to discriminate

different folding types of globular proteins. The discrimination accuracy

obtained with amino acid occurrence is better than that obtained with amino acid

composition and/or amino acid properties. In addition, the method is very fast to

obtain the results.

DOI: 10.1186/1471-2105-8-404

PMCID: PMC2174517

PMID: 17953741 [Indexed for MEDLINE]

2598. Bioinformatics. 2007 Oct 15;23(20):2795-6. Epub 2007 Aug 27.

TMpro web server and web service: transmembrane helix prediction through amino

acid property analysis.

Ganapathiraju M(1), Jursa CJ, Karimi HA, Klein-Seetharaman J.

Author information:

(1)Language Technologies Institute, School of Computer Science, Carnegie Mellon

University, Pittsburgh, PA 15213, USA.

TMpro is a transmembrane (TM) helix prediction algorithm that uses language

processing methodology for TM segment identification. It is primarily based on

the analysis of statistical distributions of properties of amino acids in

transmembrane segments. This article describes the availability of TMpro on the

internet via a web interface. The key features of the interface are: (i) output

is generated in multiple formats including a user-interactive graphical chart

which allows comparison of TMpro predicted segment locations with other labeled

segments input by the user, such as predictions from other methods. (ii) Up to

5000 sequences can be submitted at a time for prediction. (iii) TMpro is

available as a web server and is published as a web service so that the method

can be accessed by users as well as other services depending on the need for data

integration.AVAILABILITY: http://linzer.blm.cs.cmu.edu/tmpro/ (web server and

help), http://blm.sis.pitt.edu:8080/axis/services/TMProFetcherService (web

service).

DOI: 10.1093/bioinformatics/btm398

PMCID: PMC3263380

PMID: 17724062 [Indexed for MEDLINE]

2599. Int J Syst Evol Microbiol. 2007 Oct;57(Pt 10):2259-61.

EzTaxon: a web-based tool for the identification of prokaryotes based on 16S

ribosomal RNA gene sequences.

Chun J(1), Lee JH, Jung Y, Kim M, Kim S, Kim BK, Lim YW.

Author information:

(1)Interdisciplinary Program in Bioinformatics, Seoul National University, 56-1

Shillim-dong, Kwanak-gu, Seoul 151-742, Republic of Korea. jchun@snu.ac.kr

16S rRNA gene sequences have been widely used for the identification of

prokaryotes. However, the flood of sequences of non-type strains and the lack of

a peer-reviewed database for 16S rRNA gene sequences of type strains have made

routine identification of isolates difficult and labour-intensive. In the present

study, we generated a database containing 16S rRNA gene sequences of all

prokaryotic type strains. In addition, a web-based tool, named EzTaxon, for

analysis of 16S rRNA gene sequences was constructed to achieve identification of

isolates based on pairwise nucleotide similarity values and phylogenetic

inference methods. The system developed provides users with a similarity-based

search, multiple sequence alignment and various phylogenetic analyses. All of

these functions together with the 16S rRNA gene sequence database of type strains

can be successfully used for automated and reliable identification of prokaryotic

isolates. The EzTaxon server is freely accessible over the Internet at

http://www.eztaxon.org/

DOI: 10.1099/ijs.0.64915-0

PMID: 17911292 [Indexed for MEDLINE]

2600. Protein Eng Des Sel. 2007 Oct;20(10):521-3. Epub 2007 Aug 24.

The PASTA server for protein aggregation prediction.

Trovato A(1), Seno F, Tosatto SC.

Author information:

(1)Department of Physics G. Galilei, University of Padova, Padova, Italy.

Many different proteins aggregate into amyloid fibrils characterized by

cross-beta structure. beta-strands contributed by distinct protein molecules are

generally found in a parallel in-register alignment. Here, we describe the web

server for a novel algorithm, prediction of amyloid structure aggregation

(PASTA), to predict the most aggregation-prone portions and the corresponding

beta-strand inter-molecular pairing for a given input sequence. PASTA was

previously shown to yield results in excellent agreement with available

experimental observations, when tested on both natively unfolded and structured

proteins. The web server and downloadable source code are freely accessible from

the URL: http://protein.cribi.unipd.it/pasta/.

DOI: 10.1093/protein/gzm042

PMID: 17720750 [Indexed for MEDLINE]

2601. BMC Genomics. 2007 Sep 27;8:341.

Collembase: a repository for springtail genomics and soil quality assessment.

Timmermans MJ(1), de Boer ME, Nota B, de Boer TE, Mariën J, Klein-Lankhorst RM,

van Straalen NM, Roelofs D.

Author information:

(1)Vrije Universiteit, Institute of Ecological Science, Department of Animal

Ecology, De Boelelaan 1085, 1081 HV, Amsterdam, The Netherlands.

martijn.timmermans@falw.vu.nl

BACKGROUND: Environmental quality assessment is traditionally based on responses

of reproduction and survival of indicator organisms. For soil assessment the

springtail Folsomia candida (Collembola) is an accepted standard test organism.

We argue that environmental quality assessment using gene expression profiles of

indicator organisms exposed to test substrates is more sensitive, more toxicant

specific and significantly faster than current risk assessment methods. To apply

this species as a genomic model for soil quality testing we conducted an EST

sequencing project and developed an online database.

DESCRIPTION: Collembase is a web-accessible database comprising springtail (F.

candida) genomic data. Presently, the database contains information on 8686 ESTs

that are assembled into 5952 unique gene objects. Of those gene objects

approximately 40% showed homology to other protein sequences available in GenBank

(blastx analysis; non-redundant (nr) database; expect-value < 10-5). Software was

applied to infer protein sequences. The putative peptides, which had an average

length of 115 amino-acids (ranging between 23 and 440) were annotated with Gene

Ontology (GO) terms. In total 1025 peptides (approximately 17% of the gene

objects) were assigned at least one GO term (expect-value < 10-25). Within

Collembase searches can be conducted based on BLAST and GO annotation, cluster

name or using a BLAST server. The system furthermore enables easy sequence

retrieval for functional genomic and Quantitative-PCR experiments. Sequences are

submitted to GenBank (Accession numbers: EV473060 - EV481745).

CONCLUSION: Collembase http://www.collembase.org is a resource of sequence data

on the springtail F. candida. The information within the database will be linked

to a custom made microarray, based on the Agilent platform, which can be applied

for soil quality testing. In addition, Collembase supplies information that is

valuable for related scientific disciplines such as molecular ecology,

ecogenomics, molecular evolution and phylogenetics.

DOI: 10.1186/1471-2164-8-341

PMCID: PMC2234260

PMID: 17900339 [Indexed for MEDLINE]

2602. Bioinformatics. 2007 Sep 15;23(18):2504-6. Epub 2007 Aug 20.

A Laboratory Information Management System (LIMS) for a high throughput genetic

platform aimed at candidate gene mutation screening.

Voegele C(1), Tavtigian SV, de Silva D, Cuber S, Thomas A, Le Calvez-Kelm F.

Author information:

(1)International Agency for Research on Cancer, Lyon, France. voegele@iarc.fr

High throughput mutation screening in an automated environment generates large

data sets that have to be organized and stored reliably. Complex multistep

workflows require strict process management and careful data tracking. We have

developed a Laboratory Information Management Systems (LIMS) tailored to high

throughput candidate gene mutation scanning and resequencing that respects these

requirements. Designed with a client/server architecture, our system is platform

independent and based on open-source tools from the database to the web

application development strategy. Flexible, expandable and secure, the LIMS, by

communicating with most of the laboratory instruments and robots, tracks samples

and laboratory information, capturing data at every step of our automated

mutation screening workflow. An important feature of our LIMS is that it enables

tracking of information through a laboratory workflow where the process at one

step is contingent on results from a previous step.AVAILABILITY: Script for MySQL

database table creation and source code of the whole JSP application are freely

available on our website: http://www-gcs.iarc.fr/lims/.

SUPPLEMENTARY INFORMATION: System server configuration, database structure and

additional details on the LIMS and the mutation screening workflow are available

on our website: http://www-gcs.iarc.fr/lims/

DOI: 10.1093/bioinformatics/btm365

PMID: 17709339 [Indexed for MEDLINE]

2603. Bioinformatics. 2007 Sep 15;23(18):2495-7. Epub 2007 Jul 21.

APID2NET: unified interactome graphic analyzer.

Hernandez-Toro J(1), Prieto C, De las Rivas J.

Author information:

(1)Bioinformatics and Functional Genomics Research Group, Cancer Research Center

(IBMCC-CIC, CSIC-USAL), Salamanca, Spain.

MOTIVATION: Exploration and analysis of interactome networks at systems level

requires unification of the biomolecular elements and annotations that come from

many different high-throughput or small-scale proteomic experiments. Only such

integration can provide a non-redundant and consistent identification of proteins

and interactions. APID2NET is a new tool that works with Cytoscape to allow

surfing unified interactome data by querying APID server

(http://bioinfow.dep.usal.es/apid/) to provide interactive analysis of

protein-protein interaction (PPI) networks. The program is designed to visualize,

explore and analyze the proteins and interactions retrieved, including the

annotations and attributes associated to them, such as: GO terms, InterPro

domains, experimental methods that validate each interaction, PubMed IDs, UniProt

IDs, etc. The tool provides interactive graphical representation of the networks

with all Cytoscape capabilities, plus new automatic tools to find concurrent

functional and structural attributes along all protein pairs in a network.

AVAILABILITY: http://bioinfow.dep.usal.es/apid/apid2net.html.

SUPPLEMENTARY INFORMATION: Installation Guide and User's Guide are supplied at

the Web site indicated above.

DOI: 10.1093/bioinformatics/btm373

PMID: 17644818 [Indexed for MEDLINE]

2604. BMC Bioinformatics. 2007 Sep 13;8:338.

OntologyWidget - a reusable, embeddable widget for easily locating ontology

terms.

Beauheim CC(1), Wymore F, Nitzberg M, Zachariah ZK, Jin H, Skene JH, Ball CA,

Sherlock G.

Author information:

(1)Department of Genetics, Stanford University School of Medicine, Stanford, CA

94305-5120, USA. beauheim@genome.stanford.edu

BACKGROUND: Biomedical ontologies are being widely used to annotate biological

data in a computer-accessible, consistent and well-defined manner. However, due

to their size and complexity, annotating data with appropriate terms from an

ontology is often challenging for experts and non-experts alike, because there

exist few tools that allow one to quickly find relevant ontology terms to easily

populate a web form.

RESULTS: We have produced a tool, OntologyWidget, which allows users to rapidly

search for and browse ontology terms. OntologyWidget can easily be embedded in

other web-based applications. OntologyWidget is written using AJAX (Asynchronous

JavaScript and XML) and has two related elements. The first is a dynamic

auto-complete ontology search feature. As a user enters characters into the

search box, the appropriate ontology is queried remotely for terms that match the

typed-in text, and the query results populate a drop-down list with all potential

matches. Upon selection of a term from the list, the user can locate this term

within a generic and dynamic ontology browser, which comprises the second element

of the tool. The ontology browser shows the paths from a selected term to the

root as well as parent/child tree hierarchies. We have implemented web services

at the Stanford Microarray Database (SMD), which provide the OntologyWidget with

access to over 40 ontologies from the Open Biological Ontology (OBO) website 1.

Each ontology is updated weekly. Adopters of the OntologyWidget can either use

SMD's web services, or elect to rely on their own. Deploying the OntologyWidget

can be accomplished in three simple steps: (1) install Apache Tomcat 2 on one's

web server, (2) download and install the OntologyWidget servlet stub that

provides access to the SMD ontology web services, and (3) create an html

(HyperText Markup Language) file that refers to the OntologyWidget using a

simple, well-defined format.

CONCLUSION: We have developed OntologyWidget, an easy-to-use ontology search and

display tool that can be used on any web page by creating a simple html

description. OntologyWidget provides a rapid auto-complete search function paired

with an interactive tree display. We have developed a web service layer that

communicates between the web page interface and a database of ontology terms. We

currently store 40 of the ontologies from the OBO website 1, as well as a several

others. These ontologies are automatically updated on a weekly basis.

OntologyWidget can be used in any web-based application to take advantage of the

ontologies we provide via web services or any other ontology that is provided

elsewhere in the correct format. The full source code for the JavaScript and

description of the OntologyWidget is available from

http://smd.stanford.edu/ontologyWidget/.

DOI: 10.1186/1471-2105-8-338

PMCID: PMC2080642

PMID: 17854506 [Indexed for MEDLINE]

2605. BMC Bioinformatics. 2007 Sep 13;8:337.

Support Vector Machine-based method for predicting subcellular localization of

mycobacterial proteins using evolutionary information and motifs.

Rashid M(1), Saha S, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector-39A,

Chandigarh, India. mamoon@imtech.res.in

BACKGROUND: In past number of methods have been developed for predicting

subcellular location of eukaryotic, prokaryotic (Gram-negative and Gram-positive

bacteria) and human proteins but no method has been developed for mycobacterial

proteins which may represent repertoire of potent immunogens of this dreaded

pathogen. In this study, attempt has been made to develop method for predicting

subcellular location of mycobacterial proteins.

RESULTS: The models were trained and tested on 852 mycobacterial proteins and

evaluated using five-fold cross-validation technique. First SVM (Support Vector

Machine) model was developed using amino acid composition and overall accuracy of

82.51% was achieved with average accuracy (mean of class-wise accuracy) of

68.47%. In order to utilize evolutionary information, a SVM model was developed

using PSSM (Position-Specific Scoring Matrix) profiles obtained from PSI-BLAST

(Position-Specific Iterated BLAST) and overall accuracy achieved was of 86.62%

with average accuracy of 73.71%. In addition, HMM (Hidden Markov Model),

MEME/MAST (Multiple Em for Motif Elicitation/Motif Alignment and Search Tool) and

hybrid model that combined two or more models were also developed. We achieved

maximum overall accuracy of 86.8% with average accuracy of 89.00% using

combination of PSSM based SVM model and MEME/MAST. Performance of our method was

compared with that of the existing methods developed for predicting subcellular

locations of Gram-positive bacterial proteins.

CONCLUSION: A highly accurate method has been developed for predicting

subcellular location of mycobacterial proteins. This method also predicts very

important class of proteins that is membrane-attached proteins. This method will

be useful in annotating newly sequenced or hypothetical mycobacterial proteins.

Based on above study, a freely accessible web server TBpred

http://www.imtech.res.in/raghava/tbpred/ has been developed.

DOI: 10.1186/1471-2105-8-337

PMCID: PMC2147037

PMID: 17854501 [Indexed for MEDLINE]

2606. Bioinformatics. 2007 Sep 1;23(17):2334-6. Epub 2007 Jun 22.

ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple

databases and user's datasets.

Deng W(1), Nickle DC, Learn GH, Maust B, Mullins JI.

Author information:

(1)Department of Microbiology, University of Washington School of Medicine,

Seattle, WA 98195, USA.

ViroBLAST is a stand-alone BLAST web interface for nucleotide and amino acid

sequence similarity searches. It extends the utility of BLAST to query against

multiple sequence databases and user sequence datasets, and provides a friendly

output to easily parse and navigate BLAST results. ViroBLAST is readily useful

for all research areas that require BLAST functions and is available online and

as a downloadable archive for independent installation.AVAILABILITY:

http://indra.mullins.microbiol.washington.edu/blast/viroblast.php.

DOI: 10.1093/bioinformatics/btm331

PMID: 17586542 [Indexed for MEDLINE]

2607. Comput Methods Programs Biomed. 2007 Sep;87(3):230-8. Epub 2007 Jul 17.

RFRCDB-siRNA: improved design of siRNAs by random forest regression model coupled

with database searching.

Jiang P(1), Wu H, Da Y, Sang F, Wei J, Sun X, Lu Z.

Author information:

(1)State Key Laboratory of Bioelectronics, Department of Biological Science and

Medical Engineering, Southeast University, Nanjing 210096, PR China.

Although the observations concerning the factors which influence the siRNA

efficacy give clues to the mechanism of RNAi, the quantitative prediction of the

siRNA efficacy is still a challenge task. In this paper, we introduced a novel

non-linear regression method: random forest regression (RFR), to quantitatively

estimate siRNAs efficacy values. Compared with an alternative machine learning

regression algorithm, support vector machine regression (SVR) and four other

score-based algorithms [A. Reynolds, D. Leake, Q. Boese, S. Scaringe, W.S.

Marshall, A. Khvorova, Rational siRNA design for RNA interference, Nat.

Biotechnol. 22 (2004) 326-330; K. Ui-Tei, Y. Naito, F. Takahashi, T. Haraguchi,

H. Ohki-Hamazaki, A. Juni, R. Ueda, K. Saigo, Guidelines for the selection of

highly effective siRNA sequences for mammalian and chick RNA interference,

Nucleic Acids Res. 32 (2004) 936-948; A.C. Hsieh, R. Bo, J. Manola, F. Vazquez,

O. Bare, A. Khvorova, S. Scaringe, W.R. Sellers, A library of siRNA duplexes

targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing

for use in cell-based screens, Nucleic Acids Res. 32 (2004) 893-901; M.

Amarzguioui, H. Prydz, An algorithm for selection of functional siRNA sequences,

Biochem. Biophys. Res. Commun. 316 (2004) 1050-1058) our RFR model achieved the

best performance of all. A web-server, RFRCDB-siRNA

(http://www.bioinf.seu.edu.cn/siRNA/index.htm), has been developed. RFRCDB-siRNA

consists of two modules: a siRNA-centric database and a RFR prediction system.

RFRCDB-siRNA works as follows: (1) Instead of directly predicting the gene

silencing activity of siRNAs, the service takes these siRNAs as queries to search

against the siRNA-centric database. The matched sequences with the exceeding the

user defined functionality value threshold are kept. (2) The mismatched sequences

are then processed into the RFR prediction system for further analysis.

DOI: 10.1016/j.cmpb.2007.06.001

PMID: 17644215 [Indexed for MEDLINE]

2608. Med Inform Internet Med. 2007 Sep;32(3):199-214.

Knowledge-based generation of diagnostic hypotheses and therapy recommendations

for toxoplasma infections in pregnancy.

Kopecky D(1), Adlassnig KP, Prusa AR, Hayde M, Hayashi Y, Panzenböck B,

Rappelsberger A, Pollak A.

Author information:

(1)Section on Medical Expert and Knowledge-Based Systems, Core Unit for Medical

Statistics and Informatics, Medical University of Vienna, Vienna, Austria.

Primary infection of pregnant women with the parasite Toxoplasma gondii results

in infections of the unborn by transplacental transmission in about 50% of the

cases. The degree of possible damage depends on the duration of parasitical

impact on fetal tissues. The web-based software system ToxoNet processes the

results of serological antibody tests performed during pregnancy by means of a

knowledge base containing medical knowledge on the interpretation of

toxoplasmosis serology findings. For this purpose, it matches the results of all

serological investigations of maternal blood with the content of the knowledge

base and generates interpretive reports consisting of a diagnostic hypothesis,

recommendations for therapy, and proposals for further investigations. Fuzzy sets

are used to formalize certain intervals between subsequent investigations to take

the varying immune responses of individual patients into account. In a

retrospective study, ToxoNet classified 100% of the trivial serological cases and

about 87.8% of the more complex cases correctly. ToxoNet comprises a knowledge

base, a system for interpretation, and a knowledge acquisition and modification

program. It is available on the WWW by accessing a medical knowledge-base server

via standard browsers.

DOI: 10.1080/14639230701446570

PMID: 17701826 [Indexed for MEDLINE]

2609. BMC Bioinformatics. 2007 Aug 29;8:316.

Versatile annotation and publication quality visualization of protein complexes

using POLYVIEW-3D.

Porollo A(1), Meller J.

Author information:

(1)Department of Environmental Health, University of Cincinnati, Cincinnati, OH

45267, USA. alexey.porollo@uc.edu

BACKGROUND: Macromolecular visualization as well as automated structural and

functional annotation tools play an increasingly important role in the

post-genomic era, contributing significantly towards the understanding of

molecular systems and processes. For example, three dimensional (3D) models help

in exploring protein active sites and functional hot spots that can be targeted

in drug design. Automated annotation and visualization pipelines can also reveal

other functionally important attributes of macromolecules. These goals are

dependent on the availability of advanced tools that integrate better the

existing databases, annotation servers and other resources with state-of-the-art

rendering programs.

RESULTS: We present a new tool for protein structure analysis, with the focus on

annotation and visualization of protein complexes, which is an extension of our

previously developed POLYVIEW web server. By integrating the web technology with

state-of-the-art software for macromolecular visualization, such as the PyMol

program, POLYVIEW-3D enables combining versatile structural and functional

annotations with a simple web-based interface for creating publication quality

structure rendering, as well as animated images for Powerpoint, web sites and

other electronic resources. The service is platform independent and no plug-ins

are required. Several examples of how POLYVIEW-3D can be used for structural and

functional analysis in the context of protein-protein interactions are presented

to illustrate the available annotation options.

CONCLUSION: POLYVIEW-3D server features the PyMol image rendering that provides

detailed and high quality presentation of macromolecular structures, with an easy

to use web-based interface. POLYVIEW-3D also provides a wide array of options for

automated structural and functional analysis of proteins and their complexes.

Thus, the POLYVIEW-3D server may become an important resource for researches and

educators in the fields of protein science and structural bioinformatics. The new

server is available at http://polyview.cchmc.org/polyview3d.html.

DOI: 10.1186/1471-2105-8-316

PMCID: PMC1978507

PMID: 17727718 [Indexed for MEDLINE]

2610. BMC Bioinformatics. 2007 Aug 28;8:312.

MaxAlign: maximizing usable data in an alignment.

Gouveia-Oliveira R(1), Sackett PW, Pedersen AG.

Author information:

(1)Center for Biological Sequence Analysis, Technical University of Denmark,

Building 208, DK-2800 Lyngby, Denmark. rodrigo@cbs.dtu.dk

BACKGROUND: The presence of gaps in an alignment of nucleotide or protein

sequences is often an inconvenience for bioinformatical studies. In phylogenetic

and other analyses, for instance, gapped columns are often discarded entirely

from the alignment.

RESULTS: MaxAlign is a program that optimizes the alignment prior to such

analyses. Specifically, it maximizes the number of nucleotide (or amino acid)

symbols that are present in gap-free columns - the alignment area - by selecting

the optimal subset of sequences to exclude from the alignment. MaxAlign can be

used prior to phylogenetic and bioinformatical analyses as well as in other

situations where this form of alignment improvement is useful. In this work we

test MaxAlign's performance in these tasks and compare the accuracy of

phylogenetic estimates including and excluding gapped columns from the analysis,

with and without processing with MaxAlign. In this paper we also introduce a new

simple measure of tree similarity, Normalized Symmetric Similarity (NSS) that we

consider useful for comparing tree topologies.

CONCLUSION: We demonstrate how MaxAlign is helpful in detecting misaligned or

defective sequences without requiring manual inspection. We also show that it is

not advisable to exclude gapped columns from phylogenetic analyses unless

MaxAlign is used first. Finally, we find that the sequences removed by MaxAlign

from an alignment tend to be those that would otherwise be associated with low

phylogenetic accuracy, and that the presence of gaps in any given sequence does

not seem to disturb the phylogenetic estimates of other sequences. The MaxAlign

web-server is freely available online at http://www.cbs.dtu.dk/services/MaxAlign

where supplementary information can also be found. The program is also freely

available as a Perl stand-alone package.

DOI: 10.1186/1471-2105-8-312

PMCID: PMC2000915

PMID: 17725821 [Indexed for MEDLINE]

2611. Biochem Biophys Res Commun. 2007 Aug 24;360(2):339-45. Epub 2007 Jun 15.

MemType-2L: a web server for predicting membrane proteins and their types by

incorporating evolution information through Pse-PSSM.

Chou KC(1), Shen HB.

Author information:

(1)Gordon Life Science Institute, San Diego, CA 92130, USA.

kcchou@gordonlifescience.org

Given an uncharacterized protein sequence, how can we identify whether it is a

membrane protein or not? If it is, which membrane protein type it belongs to?

These questions are important because they are closely relevant to the biological

function of the query protein and to its interaction process with other molecules

in a biological system. Particularly, with the avalanche of protein sequences

generated in the Post-Genomic Age and the relatively much slower progress in

using biochemical experiments to determine their functions, it is highly desired

to develop an automated method that can be used to help address these questions.

In this study, a 2-layer predictor, called MemType-2L, has been developed: the

1st layer prediction engine is to identify a query protein as membrane or

non-membrane; if it is a membrane protein, the process will be automatically

continued with the 2nd-layer prediction engine to further identify its type among

the following eight categories: (1) type I, (2) type II, (3) type III, (4) type

IV, (5) multipass, (6) lipid-chain-anchored, (7) GPI-anchored, and (8)

peripheral. MemType-2L is featured by incorporating the evolution information

through representing the protein samples with the Pse-PSSM (Pseudo

Position-Specific Score Matrix) vectors, and by containing an ensemble classifier

formed by fusing many powerful individual OET-KNN (Optimized Evidence-Theoretic

K-Nearest Neighbor) classifiers. The success rates obtained by MemType-2L on a

new-constructed stringent dataset by both the jackknife test and the independent

dataset test are quite high, indicating that MemType-2L may become a very useful

high throughput tool. As a Web server, MemType-2L is freely accessible to the

public at http://chou.med.harvard.edu/bioinf/MemType.

DOI: 10.1016/j.bbrc.2007.06.027

PMID: 17586467 [Indexed for MEDLINE]

2612. BMC Bioinformatics. 2007 Aug 21;8:304.

Evaluation of 3D-Jury on CASP7 models.

Kaján L(1), Rychlewski L.

Author information:

(1)BioInfoBank Institute, ul, Limanowskiego 24 A, 60-744 Poznań, Poland.

kajla@bioinfo.pl

BACKGROUND: 3D-Jury, the structure prediction consensus method publicly available

in the Meta Server http://meta.bioinfo.pl/, was evaluated using models gathered

in the 7th round of the Critical Assessment of Techniques for Protein Structure

Prediction (CASP7). 3D-Jury is an automated expert process that generates protein

structure meta-predictions from sets of models obtained from partner servers.

RESULTS: The performance of 3D-Jury was analysed for three aspects. First, we

examined the correlation between the 3D-Jury score and a model quality measure:

the number of correctly predicted residues. The 3D-Jury score was shown to

correlate significantly with the number of correctly predicted residues, the

correlation is good enough to be used for prediction. 3D-Jury was also found to

improve upon the competing servers' choice of the best structure model in most

cases. The value of the 3D-Jury score as a generic reliability measure was also

examined. We found that the 3D-Jury score separates bad models from good models

better than the reliability score of the original server in 27 cases and falls

short of it in only 5 cases out of a total of 38. We report the release of a new

Meta Server feature: instant 3D-Jury scoring of uploaded user models.

CONCLUSION: The 3D-Jury score continues to be a good indicator of structural

model quality. It also provides a generic reliability score, especially important

for models that were not assigned such by the original server. Individual

structure modellers can also benefit from the 3D-Jury scoring system by testing

their models in the new instant scoring feature

http://meta.bioinfo.pl/compare\_your\_model\_example.pl available in the Meta

Server.

DOI: 10.1186/1471-2105-8-304

PMCID: PMC2040163

PMID: 17711571 [Indexed for MEDLINE]

2613. BMC Bioinformatics. 2007 Aug 20;8:302.

Prediction of the burial status of transmembrane residues of helical membrane

proteins.

Park Y(1), Hayat S, Helms V.

Author information:

(1)Center for Bioinformatics, Saarland University, 66041 Saarbruecken, Germany.

yungki@bioinformatik.uni-saarland.de

BACKGROUND: Helical membrane proteins (HMPs) play a crucial role in diverse

cellular processes, yet it still remains extremely difficult to determine their

structures by experimental techniques. Given this situation, it is highly

desirable to develop sequence-based computational methods for predicting

structural characteristics of HMPs.

RESULTS: We have developed TMX (TransMembrane eXposure), a novel method for

predicting the burial status (i.e. buried in the protein structure vs. exposed to

the membrane) of transmembrane (TM) residues of HMPs. TMX derives positional

scores of TM residues based on their profiles and conservation indices. Then, a

support vector classifier is used for predicting their burial status. Its

prediction accuracy is 78.71% on a benchmark data set, representing considerable

improvements over 68.67% and 71.06% of previously proposed methods. Importantly,

unlike the previous methods, TMX automatically yields confidence scores for the

predictions made. In addition, a feature selection incorporated in TMX reveals

interesting insights into the structural organization of HMPs.

CONCLUSION: A novel computational method, TMX, has been developed for predicting

the burial status of TM residues of HMPs. Its prediction accuracy is much higher

than that of previously proposed methods. It will be useful in elucidating

structural characteristics of HMPs as an inexpensive, auxiliary tool. A web

server for TMX is established at http://service.bioinformatik.uni-saarland.de/tmx

and freely available to academic users, along with the data set used.

DOI: 10.1186/1471-2105-8-302

PMCID: PMC2000914

PMID: 17708758 [Indexed for MEDLINE]

2614. Proteins. 2007 Aug 15;68(3):636-45.

Fold recognition by concurrent use of solvent accessibility and residue depth.

Liu S(1), Zhang C, Liang S, Zhou Y.

Author information:

(1)Howard Hughes Medical Institute Center for Single Molecule Biophysics,

Department of Physiology and Biophysics, State University of New York at Buffalo,

Buffalo, New York 14214, USA.

Recognizing the structural similarity without significant sequence identity

(called fold recognition) is the key for bridging the gap between the number of

known protein sequences and the number of structures solved. Previously, we

developed a fold-recognition method called SP(3) which combines sequence-derived

sequence profiles, secondary-structure profiles and residue-depth dependent,

structure-derived sequence profiles. The use of residue-depth-dependent profiles

makes SP(3) one of the best automatic predictors in CASP 6. Because residue depth

(RD) and solvent accessible surface area (solvent accessibility) are

complementary in describing the exposure of a residue to solvent, we test whether

or not incorporation of solvent-accessibility profiles into SP(3) could further

increase the accuracy of fold recognition. The resulting method, called SP(4),

was tested in SALIGN benchmark for alignment accuracy and Lindahl, LiveBench 8

and CASP7 blind prediction for fold recognition sensitivity and model-structure

accuracy. For remote homologs, SP(4) is found to consistently improve over SP(3)

in the accuracy of sequence alignment and predicted structural models as well as

in the sensitivity of fold recognition. Our result suggests that RD and solvent

accessibility can be used concurrently for improving the accuracy and sensitivity

of fold recognition. The SP(4) server and its local usage package are available

on http://sparks.informatics.iupui.edu/SP4.

DOI: 10.1002/prot.21459

PMID: 17510969 [Indexed for MEDLINE]

2615. Acta Crystallogr D Biol Crystallogr. 2007 Aug;63(Pt 8):935-8. Epub 2007 Jul 17.

ValLigURL: a server for ligand-structure comparison and validation.

Kleywegt GJ(1), Harris MR.

Author information:

(1)Department of Cell and Molecular Biology, Uppsala University, Biomedical

Centre, Box 596, SE-751 24 Uppsala, Sweden. gerard@xray.bmc.uu.se

A new web-based tool called ValLigURL is described. It can be used by practising

crystallographers to validate the geometry of a ligand and to compare the

conformation of a ligand with all instances of that ligand in the structural

database (wwPDB). In addition, it can be used by structural bioinformaticians to

survey the quality or conformational diversity of any ligand across the entire

structural database. The server is freely accessible at the URL

http://eds.bmc.uu.se/eds/valligurl.php.

DOI: 10.1107/S090744490703315X

PMID: 17642521 [Indexed for MEDLINE]

2616. PLoS Comput Biol. 2007 Aug;3(8):e160.

Automated protein subfamily identification and classification.

Brown DP(1), Krishnamurthy N, Sjölander K.

Author information:

(1)Department of Bioengineering, University of California, Berkeley, California,

United States of America.

Function prediction by homology is widely used to provide preliminary functional

annotations for genes for which experimental evidence of function is unavailable

or limited. This approach has been shown to be prone to systematic error,

including percolation of annotation errors through sequence databases.

Phylogenomic analysis avoids these errors in function prediction but has been

difficult to automate for high-throughput application. To address this

limitation, we present a computationally efficient pipeline for phylogenomic

classification of proteins. This pipeline uses the SCI-PHY (Subfamily

Classification in Phylogenomics) algorithm for automatic subfamily

identification, followed by subfamily hidden Markov model (HMM) construction. A

simple and computationally efficient scoring scheme using family and subfamily

HMMs enables classification of novel sequences to protein families and

subfamilies. Sequences representing entirely novel subfamilies are differentiated

from those that can be classified to subfamilies in the input training set using

logistic regression. Subfamily HMM parameters are estimated using an

information-sharing protocol, enabling subfamilies containing even a single

sequence to benefit from conservation patterns defining the family as a whole or

in related subfamilies. SCI-PHY subfamilies correspond closely to functional

subtypes defined by experts and to conserved clades found by phylogenetic

analysis. Extensive comparisons of subfamily and family HMM performances show

that subfamily HMMs dramatically improve the separation between homologous and

non-homologous proteins in sequence database searches. Subfamily HMMs also

provide extremely high specificity of classification and can be used to predict

entirely novel subtypes. The SCI-PHY Web server at

http://phylogenomics.berkeley.edu/SCI-PHY/ allows users to upload a multiple

sequence alignment for subfamily identification and subfamily HMM construction.

Biologists wishing to provide their own subfamily definitions can do so. Source

code is available on the Web page. The Berkeley Phylogenomics Group PhyloFacts

resource contains pre-calculated subfamily predictions and subfamily HMMs for

more than 40,000 protein families and domains at

http://phylogenomics.berkeley.edu/phylofacts/.

DOI: 10.1371/journal.pcbi.0030160

PMCID: PMC1950344

PMID: 17708678 [Indexed for MEDLINE]

2617. Protein Sci. 2007 Aug;16(8):1569-76.

Toward rational protein crystallization: A Web server for the design of

crystallizable protein variants.

Goldschmidt L(1), Cooper DR, Derewenda ZS, Eisenberg D.

Author information:

(1)Howard Hughes Medical Institute, University of California, Los Angeles-DOE

Institute of Genomics and Proteomics, Los Angeles, California 90095-1570, USA.

Growing well-diffracting crystals constitutes a serious bottleneck in structural

biology. A recently proposed crystallization methodology for "stubborn

crystallizers" is to engineer surface sequence variants designed to form

intermolecular contacts that could support a crystal lattice. This approach

relies on the concept of surface entropy reduction (SER), i.e., the replacement

of clusters of flexible, solvent-exposed residues with residues with lower

conformational entropy. This strategy minimizes the loss of conformational

entropy upon crystallization and renders crystallization thermodynamically

favorable. The method has been successfully used to crystallize more than 15

novel proteins, all stubborn crystallizers. But the choice of suitable sites for

mutagenesis is not trivial. Herein, we announce a Web server, the surface entropy

reduction prediction server (SERp server), designed to identify mutations that

may facilitate crystallization. Suggested mutations are predicted based on an

algorithm incorporating a conformational entropy profile, a secondary structure

prediction, and sequence conservation. Minor considerations include the nature of

flanking residues and gaps between mutation candidates. While designed to be used

with default values, the server has many user-controlled parameters allowing for

considerable flexibility. Within, we discuss (1) the methodology of the server,

(2) how to interpret the results, and (3) factors that must be considered when

selecting mutations. We also attempt to benchmark the server by comparing the

server's predictions with successful SER structures. In most cases, the structure

yielding mutations were easily identified by the SERp server. The server can be

accessed at http://www.doe-mbi.ucla.edu/Services/SER.

DOI: 10.1110/ps.072914007

PMCID: PMC2203352

PMID: 17656576 [Indexed for MEDLINE]

2618. Proteomics. 2007 Aug;7(15):2553-6.

myProMS, a web server for management and validation of mass spectrometry-based

proteomic data.

Poullet P(1), Carpentier S, Barillot E.

Author information:

(1)Institut Curie, Service de Bioinformatique, Paris, France.

patrick.poullet@curie.fr

Curation and interpretation of protein databank-search results by human experts

are key aspects of MS-based proteomic data acquisition. These tasks are often

overlooked due to the vast amount of data to inspect. We have developed myProMS,

a web server designed to ease search results validation and interpretation by

improving data organization, mining and sharing between MS specialists and

biologists during MS-based collaborative projects. A demo is accessible at

http://bioinfo.curie.fr/myproms.

DOI: 10.1002/pmic.200600784

PMID: 17610305 [Indexed for MEDLINE]

2619. Yi Chuan. 2007 Aug;29(8):1023-6.

[GSDS: a gene structure display server].

[Article in Chinese]

Guo AY(1), Zhu QH, Chen X, Luo JC.

Author information:

(1)Center for Bioinformatics, Peking University, Beijing 100871, China.

guoay@mail.cbi.pku.edu.cn

We developed a web server GSDS (Gene Structure Display Server) for drawing gene

structure schematic diagrams. Users can submit three types of dataCDS and genomic

sequences, NCBI GenBank accession numbers or GIs, exon positions on a gene. GSDS

uses this information to obtain the gene structure and draw diagram for it. Users

can also designate some special regions to mark on the gene structure diagram.

The output result will be PNG or SVG format picture. The corresponding sequence

will be shown in a new window by clicking the picture in PNG format. A Chinese

version for the main page is also built. The GSDS is available on

http://gsds.cbi.pku.edu.cn/.

PMID: 17681935 [Indexed for MEDLINE]

2620. BMC Bioinformatics. 2007 Jul 25;8:266.

BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies

based on overlapping genes.

Luo Y(1), Fu C, Zhang DY, Lin K.

Author information:

(1)MOE Key Laboratory for Biodiversity Science and Ecological Engineering and

College of Life Sciences, Beijing Normal University, Beijing, China.

luoyq@mail.bnu.edu.cn <luoyq@mail.bnu.edu.cn>

BACKGROUND: Overlapping genes (OGs) in bacterial genomes are pairs of adjacent

genes of which the coding sequences overlap partly or entirely. With the rapid

accumulation of sequence data, many OGs in bacterial genomes have now been

identified. Indeed, these might prove a consistent feature across all microbial

genomes. Our previous work suggests that OGs can be considered as robust markers

at the whole genome level for the construction of phylogenies. An online,

interactive web server for inferring phylogenies is needed for biologists to

analyze phylogenetic relationships among a set of bacterial genomes of interest.

DESCRIPTION: BPhyOG is an online interactive server for reconstructing the

phylogenies of completely sequenced bacterial genomes on the basis of their

shared overlapping genes. It provides two tree-reconstruction methods: Neighbor

Joining (NJ) and Unweighted Pair-Group Method using Arithmetic averages (UPGMA).

Users can apply the desired method to generate phylogenetic trees, which are

based on an evolutionary distance matrix for the selected genomes. The distance

between two genomes is defined by the normalized number of their shared OG pairs.

BPhyOG also allows users to browse the OGs that were used to infer the

phylogenetic relationships. It provides detailed annotation for each OG pair and

the features of the component genes through hyperlinks. Users can also retrieve

each of the homologous OG pairs that have been determined among 177 genomes. It

is a useful tool for analyzing the tree of life and overlapping genes from a

genomic standpoint.

CONCLUSION: BPhyOG is a useful interactive web server for genome-wide inference

of any potential evolutionary relationship among the genomes selected by users.

It currently includes 177 completely sequenced bacterial genomes containing

79,855 OG pairs, the annotation and homologous OG pairs of which are integrated

comprehensively. The reliability of phylogenies complemented by annotations make

BPhyOG a powerful web server for genomic and genetic studies. It is freely

available at http://cmb.bnu.edu.cn/BPhyOG.

DOI: 10.1186/1471-2105-8-266

PMCID: PMC1940028

PMID: 17650344 [Indexed for MEDLINE]

2621. BMC Genomics. 2007 Jul 24;8:246.

EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool

for agronomical species.

Zhou X(1), Su Z.

Author information:

(1)Division of Bioinformatics, State Key Laboratory of Plant Physiology and

Biochemistry, College of Biological Sciences, China Agricultural University,

Beijing 100094, China. xzhou82@gmail.com

BACKGROUND: It is always difficult to interpret microarray results. Recently, a

handful of tools have been developed to meet this need, but almost none of them

were designed to support agronomical species.

DESCRIPTION: This paper presents EasyGO, a web server to perform Gene Ontology

based functional interpretation on groups of genes or GeneChip probe sets. EasyGO

makes a special contribution to the agronomical research community by supporting

Affymetrix GeneChips of both crops and farm animals and by providing stronger

capabilities for results visualization and user interaction. Currently it

supports 11 agronomical plants, 3 farm animals, and the model plant Arabidopsis.

The authors demonstrated EasyGO's ability to uncover hidden knowledge by

analyzing a group of probe sets with similar expression profiles.

CONCLUSION: EasyGO is a good tool for helping biologists and agricultural

scientists to discover enriched biological knowledge that can provide solutions

or suggestions for original problems. It is freely available to all users at

http://bioinformatics.cau.edu.cn/easygo/.

DOI: 10.1186/1471-2164-8-246

PMCID: PMC1940007

PMID: 17645808 [Indexed for MEDLINE]

2622. BMC Bioinformatics. 2007 Jul 23;8:263.

Analysis and prediction of antibacterial peptides.

Lata S(1), Sharma BK, Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector39A, Chandigarh, India.

sneh@imtech.res.in

BACKGROUND: Antibacterial peptides are important components of the innate immune

system, used by the host to protect itself from different types of pathogenic

bacteria. Over the last few decades, the search for new drugs and drug targets

has prompted an interest in these antibacterial peptides. We analyzed 486

antibacterial peptides, obtained from antimicrobial peptide database APD, in

order to understand the preference of amino acid residues at specific positions

in these peptides.

RESULTS: It was observed that certain types of residues are preferred over others

in antibacterial peptides, particularly at the N and C terminus. These

observations encouraged us to develop a method for predicting antibacterial

peptides in proteins from their amino acid sequence. First, the N-terminal

residues were used for predicting antibacterial peptides using Artificial Neural

Network (ANN), Quantitative Matrices (QM) and Support Vector Machine (SVM), which

resulted in an accuracy of 83.63%, 84.78% and 87.85%, respectively. Then, the

C-terminal residues were used for developing prediction methods, which resulted

in an accuracy of 77.34%, 82.03% and 85.16% using ANN, QM and SVM, respectively.

Finally, ANN, QM and SVM models were developed using N and C terminal residues,

which achieved an accuracy of 88.17%, 90.37% and 92.11%, respectively. All the

models developed in this study were evaluated using five-fold cross validation

technique. These models were also tested on an independent or blind dataset.

CONCLUSION: Among antibacterial peptides, there is preference for certain

residues at N and C termini, which helps to demarcate them from non-antibacterial

peptides. Both the termini play a crucial role in imparting the antibacterial

property to these peptides. Among the methods developed, SVM shows the best

performance in predicting antibacterial peptides followed by QM and ANN, in that

order. AntiBP (Antibacterial peptides) will help in discovering efficacious

antibacterial peptides, which we hope will prove to be a boon to combat the

dreadful antibiotic resistant bacteria. A user friendly web server has also been

developed to help the biological community, which is accessible at

http://www.imtech.res.in/raghava/antibp/.

DOI: 10.1186/1471-2105-8-263

PMCID: PMC2041956

PMID: 17645800 [Indexed for MEDLINE]

2623. BMC Plant Biol. 2007 Jul 23;7:39.

Arabidopsis Gene Family Profiler (aGFP)--user-oriented transcriptomic database

with easy-to-use graphic interface.

Dupl'áková N(1), Renák D, Hovanec P, Honysová B, Twell D, Honys D.

Author information:

(1)Laboratory of Pollen Biology, Institute of Experimental Botany ASCR v.v.i.,

Rozvojová 263, Prague 6, Czech Republic. duplakova@ueb.cas.cz

BACKGROUND: Microarray technologies now belong to the standard functional

genomics toolbox and have undergone massive development leading to increased

genome coverage, accuracy and reliability. The number of experiments exploiting

microarray technology has markedly increased in recent years. In parallel with

the rapid accumulation of transcriptomic data, on-line analysis tools are being

introduced to simplify their use. Global statistical data analysis methods

contribute to the development of overall concepts about gene expression patterns

and to query and compose working hypotheses. More recently, these applications

are being supplemented with more specialized products offering visualization and

specific data mining tools. We present a curated gene family-oriented gene

expression database, Arabidopsis Gene Family Profiler (aGFP;

http://agfp.ueb.cas.cz), which gives the user access to a large collection of

normalised Affymetrix ATH1 microarray datasets. The database currently contains

NASC Array and AtGenExpress transcriptomic datasets for various tissues at

different developmental stages of wild type plants gathered from nearly 350 gene

chips.

RESULTS: The Arabidopsis GFP database has been designed as an easy-to-use tool

for users needing an easily accessible resource for expression data of single

genes, pre-defined gene families or custom gene sets, with the further

possibility of keyword search. Arabidopsis Gene Family Profiler presents a

user-friendly web interface using both graphic and text output. Data are stored

at the MySQL server and individual queries are created in PHP script. The most

distinguishable features of Arabidopsis Gene Family Profiler database are: 1) the

presentation of normalized datasets (Affymetrix MAS algorithm and calculation of

model-based gene-expression values based on the Perfect Match-only model); 2) the

choice between two different normalization algorithms (Affymetrix MAS4 or MAS5

algorithms); 3) an intuitive interface; 4) an interactive "virtual plant"

visualizing the spatial and developmental expression profiles of both gene

families and individual genes.

CONCLUSION: Arabidopsis GFP gives users the possibility to analyze current

Arabidopsis developmental transcriptomic data starting with simple global queries

that can be expanded and further refined to visualize comparative and highly

selective gene expression profiles.

DOI: 10.1186/1471-2229-7-39

PMCID: PMC1963329

PMID: 17645793 [Indexed for MEDLINE]

2624. Bioinformatics. 2007 Jul 15;23(14):1843-5. Epub 2007 May 7.

CRCView: a web server for analyzing and visualizing microarray gene expression

data using model-based clustering.

Xiang Z(1), Qin ZS, He Y.

Author information:

(1)Unit for Laboratory Animal Medicine, University of Michigan, Ann Arbor, MI,

USA.

CRCView is a user-friendly point-and-click web server for analyzing and

visualizing microarray gene expression data using a Dirichlet process mixture

model-based clustering algorithm. CRCView is designed to clustering genes based

on their expression profiles. It allows flexible input data format, rich

graphical illustration as well as integrated GO term based

annotation/interpretation of clustering results.AVAILABILITY:

http://helab.bioinformatics.med.umich.edu/crcview/.

DOI: 10.1093/bioinformatics/btm238

PMID: 17485426 [Indexed for MEDLINE]

2625. BMC Genomics. 2007 Jul 10;8:225.

EuMicroSatdb: a database for microsatellites in the sequenced genomes of

eukaryotes.

Aishwarya V(1), Grover A, Sharma PC.

Author information:

(1)University School of Biotechnology, Guru Gobind Singh Indraprastha University,

Kashmere Gate, Delhi, India. veenu\_aishwarya@yahoo.com

<veenu\_aishwarya@yahoo.com>

BACKGROUND: Microsatellites have immense utility as molecular markers in

different fields like genome characterization and mapping, phylogeny and

evolutionary biology. Existing microsatellite databases are of limited utility

for experimental and computational biologists with regard to their content and

information output. EuMicroSatdb (Eukaryotic MicroSatellite database)

http://ipu.ac.in/usbt/EuMicroSatdb.htm is a web based relational database for

easy and efficient positional mining of microsatellites from sequenced eukaryotic

genomes.

DESCRIPTION: A user friendly web interface has been developed for microsatellite

data retrieval using Active Server Pages (ASP). The backend database codes for

data extraction and assembly have been written using Perl based scripts and C++.

Precise need based microsatellites data retrieval is possible using different

input parameters like microsatellite type (simple perfect or compound perfect),

repeat unit length (mono- to hexa-nucleotide), repeat number, microsatellite

length and chromosomal location in the genome. Furthermore, information about

clustering of different microsatellites in the genome can also be retrieved.

Finally, to facilitate primer designing for PCR amplification of any desired

microsatellite locus, 200 bp upstream and downstream sequences are provided.

CONCLUSION: The database allows easy systematic retrieval of comprehensive

information about simple and compound microsatellites, microsatellite clusters

and their locus coordinates in 31 sequenced eukaryotic genomes. The information

content of the database is useful in different areas of research like gene

tagging, genome mapping, population genetics, germplasm characterization and in

understanding microsatellite dynamics in eukaryotic genomes.

DOI: 10.1186/1471-2164-8-225

PMCID: PMC1933429

PMID: 17623061 [Indexed for MEDLINE]

2626. Amino Acids. 2007 Jul;33(1):57-67. Epub 2007 Jan 19.

Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular

location prediction.

Shen HB(1), Yang J, Chou KC.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, Shanghai, China.

With the avalanche of newly-found protein sequences emerging in the post genomic

era, it is highly desirable to develop an automated method for fast and reliably

identifying their subcellular locations because knowledge thus obtained can

provide key clues for revealing their functions and understanding how they

interact with each other in cellular networking. However, predicting subcellular

location of eukaryotic proteins is a challenging problem, particularly when

unknown query proteins do not have significant homology to proteins of known

subcellular locations and when more locations need to be covered. To cope with

the challenge, protein samples are formulated by hybridizing the information

derived from the gene ontology database and amphiphilic pseudo amino acid

composition. Based on such a representation, a novel ensemble hybridization

classifier was developed by fusing many basic individual classifiers through a

voting system. Each of these basic classifiers was engineered by the KNN

(K-Nearest Neighbor) principle. As a demonstration, a new benchmark dataset was

constructed that covers the following 18 localizations: (1) cell wall, (2)

centriole, (3) chloroplast, (4) cyanelle, (5) cytoplasm, (6) cytoskeleton, (7)

endoplasmic reticulum, (8) extracell, (9) Golgi apparatus, (10) hydrogenosome,

(11) lysosome, (12) mitochondria, (13) nucleus, (14) peroxisome, (15) plasma

membrane, (16) plastid, (17) spindle pole body, and (18) vacuole. To avoid the

homology bias, none of the proteins included has > or =25% sequence identity to

any other in a same subcellular location. The overall success rates thus obtained

via the 5-fold and jackknife cross-validation tests were 81.6 and 80.3%,

respectively, which were 40-50% higher than those performed by the other existing

methods on the same strict dataset. The powerful predictor, named "Euk-PLoc", is

available as a web-server at http://202.120.37.186/bioinf/euk . Furthermore, to

support the need of people working in the relevant areas, a downloadable file

will be provided at the same website to list the results predicted by Euk-PLoc

for all eukaryotic protein entries (excluding fragments) in Swiss-Prot database

that do not have subcellular location annotations or are annotated as being

uncertain. The large-scale results will be updated twice a year to include the

new entries of eukaryotic proteins and reflect the continuous development of

Euk-PLoc.

DOI: 10.1007/s00726-006-0478-8

PMID: 17235453 [Indexed for MEDLINE]

2627. Bioinformatics. 2007 Jul 1;23(13):i539-48.

Kinetics analysis methods for approximate folding landscapes.

Tapia L(1), Tang X, Thomas S, Amato NM.

Author information:

(1)Parasol Lab, Department of Computer Science, Texas A&M University, College

Station, TX 77843, USA.

MOTIVATION: Protein motions play an essential role in many biochemical processes.

Lab studies often quantify these motions in terms of their kinetics such as the

speed at which a protein folds or the population of certain interesting states

like the native state. Kinetic metrics give quantifiable measurements of the

folding process that can be compared across a group of proteins such as a

wild-type protein and its mutants.

RESULTS: We present two new techniques, map-based master equation solution and

map-based Monte Carlo simulation, to study protein kinetics through folding rates

and population kinetics from approximate folding landscapes, models called maps.

From these two new techniques, interesting metrics that describe the folding

process, such as reaction coordinates, can also be studied. In this article we

focus on two metrics, formation of helices and structure formation around

tryptophan residues. These two metrics are often studied in the lab through

circular dichroism (CD) spectra analysis and tryptophan fluorescence experiments,

respectively. The approximated landscape models we use here are the maps of

protein conformations and their associated transitions that we have presented and

validated previously. In contrast to other methods such as the traditional master

equation and Monte Carlo simulation, our techniques are both fast and can easily

be computed for full-length detailed protein models. We validate our map-based

kinetics techniques by comparing folding rates to known experimental results. We

also look in depth at the population kinetics, helix formation and structure near

tryptophan residues for a variety of proteins.

AVAILABILITY: We invite the community to help us enrich our publicly available

database of motions and kinetics analysis by submitting to our server:

http://parasol.tamu.edu/foldingserver/.

DOI: 10.1093/bioinformatics/btm199

PMID: 17646341 [Indexed for MEDLINE]

2628. Bioinformatics. 2007 Jul 1;23(13):i392-400.

Locomotif: from graphical motif description to RNA motif search.

Reeder J(1), Reeder J, Giegerich R.

Author information:

(1)Faculty of Technology, Bielefeld University, D-33615 Bielefeld, Germany.

MOTIVATION AND RESULTS: Motivated by the recent rise of interest in small

regulatory RNAs, we present Locomotif--a new approach for locating RNA motifs

that goes beyond the previous ones in three ways: (1) motif search is based on

efficient dynamic programming algorithms, incorporating the established

thermodynamic model of RNA secondary structure formation. (2) motifs are

described graphically, using a Java-based editor, and search algorithms are

derived from the graphics in a fully automatic way. The editor allows us to draw

secondary structures, annotated with size and sequence information. They closely

resemble the established, but informal way in which RNA motifs are communicated

in the literature. Thus, the learning effort for Locomotif users is minimal. (3)

Locomotif employs a client-server approach. Motifs are designed by the user

locally. Search programs are generated and compiled on a bioinformatics server.

They are made available both for execution on the server, and for download as C

source code plus an appropriate makefile.

AVAILABILITY: Locomotif is available at

http://bibiserv.techfak.uni-bielefeld.de/locomotif.

DOI: 10.1093/bioinformatics/btm179

PMID: 17646322 [Indexed for MEDLINE]

2629. Bioinformatics. 2007 Jul 1;23(13):i175-84.

Anisotropic fluctuations of amino acids in protein structures: insights from

X-ray crystallography and elastic network models.

Eyal E(1), Chennubhotla C, Yang LW, Bahar I.

Author information:

(1)Department of Computational Biology, School of Medicine, University of

Pittsburgh, Suite 3064, Biomedical Science Tower 3, 3051 Fifth Ave., Pittsburgh,

PA 15213, USA.

MOTIVATION: A common practice in X-ray crystallographic structure refinement has

been to model atomic displacements or thermal fluctuations as isotropic motions.

Recent high-resolution data reveal, however, significant departures from

isotropy, described by anisotropic displacement parameters (ADPs) modeled for

individual atoms. Yet, ADPs are currently reported for a limited set of

structures, only.

RESULTS: We present a comparative analysis of the experimentally reported ADPs

and those theoretically predicted by the anisotropic network model (ANM) for a

representative set of structures. The relative sizes of fluctuations along

different directions are shown to agree well between experiments and theory,

while the cross-correlations between the (x-, y- and z-) components of the

fluctuations show considerable deviations. Secondary structure elements and

protein cores exhibit more robust anisotropic characteristics compared to

disordered or flexible regions. The deviations between experimental and

theoretical data are comparable to those between sets of experimental ADPs

reported for the same protein in different crystal forms. These results draw

attention to the effects of crystal form and refinement procedure on experimental

ADPs and highlight the potential utility of ANM calculations for consolidating

experimental data or assessing ADPs in the absence of experimental data.

AVAILABILITY: The ANM server at http://www.ccbb.pitt.edu/anm is upgraded to

permit users to compute and visualize the theoretical ADPs for any PDB structure,

thus providing insights into the anisotropic motions intrinsically preferred by

equilibrium structures.

SUPPLEMENTARY INFORMATION: Two Supplementary Material files can be accessed at

the journal website. The first presents the tabulated results from computations

(Pearson correlations and KL distances with respect to experimental ADPs)

reported for each of the 93 proteins in Set I (the averages over all proteins are

presented above in Table 3). The second file consists of three sections: (A)

detailed derivation of Equation (7), (B) analysis of the effect of ANM parameters

on computed ADPs and identification of parameters that achieve optimal

correlation with experiments and (C) description of the method for computing the

tangential and radial components of equilibrium fluctuations.

DOI: 10.1093/bioinformatics/btm186

PMID: 17646294 [Indexed for MEDLINE]

2630. Hum Mutat. 2007 Jul;28(7):654-9.

PKDB: Polycystic Kidney Disease Mutation Database--a gene variant database for

autosomal dominant polycystic kidney disease.

Gout AM(1), Martin NC, Brown AF, Ravine D.

Author information:

(1)The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria,

Australia.

Autosomal dominant polycystic kidney disease (ADPKD) arises from mutations in the

PKD1 and PKD2 genes. The Polycystic Kidney Disease Mutation Database (PKDB) is an

internet-accessible relational database containing comprehensive information

about germline and somatic disease-causing variants within these two genes, as

well as polymorphisms and variants of indeterminate pathogenicity. The PKDB

database structure incorporates an interface between these gene variant data and

any associated patient clinical data. An initiative of the Polycystic Kidney

Disease Foundation, PKDB is a publicly accessible database that aims to

streamline the evaluation of PKD1 and PKD2 gene variants detected in samples from

those with ADPKD, as well as to assist ongoing clinical and molecular research in

the field. As the accurate reporting of nucleotide variants is essential for

ensuring the quality of data within PKDB, a mutation checker has been mounted on

the PKDB server allowing contributors to assess the accuracy of their PKD1 and

PKD2 variant reports. Researchers and clinicians may submit their PKD1/PKD2 gene

variants and any associated deidentified clinical data via standardized

downloadable data entry forms accessible through the PKDB site. PKDB has been

launched with the full details of PKD1 and PKD2 gene variant reports published in

73 peer-reviewed articles. Through a series of user-friendly advanced search

facilities, users are able to query the database as required. The PKDB server is

accessible at http://pkdb.mayo.edu.

(c) 2007 Wiley-Liss, Inc.

DOI: 10.1002/humu.20474

PMID: 17370309 [Indexed for MEDLINE]

2631. Infect Genet Evol. 2007 Jul;7(4):463-8. Epub 2007 Jan 26.

genoBASE pylori: a genotype search tool and database of the human gastric

pathogen Helicobacter pylori.

Ahmed N(1), Majeed AA, Ahmed I, Hussain MA, Alvi A, Devi SM, Rizwan M, Ranjan A,

Sechi LA, Mégraud F.

Author information:

(1)Pathogen Evolution Group, Centre for DNA Fingerprinting and Diagnostics,

Hyderabad, India. niyaz@cdfd.org.in

Helicobacter pylori is the pathogenic bacterium linked to gastric and duodenal

ulcers and gastric carcinoma. Genomic diversity of the organism has enabled new

insights into its population biology through comparative genomics. genoBASE

pylori is an online databank of several virulence-linked and phylogenetic markers

of H. pylori strains obtained from different human populations. This

knowledgebase is built upon a relational database management system which is

connected to visualize the presence of known, pathogenicity markers such as the

co-ordinates within the cag pathogenicity island (cagPAI), the cagA gene and

motifs surrounding it, the vacA allotypes and the oipA gene frame status,

together with genotypic details in the form of DNA profiling traces and candidate

gene sequences for individual strains. This flexible search tool allows

inter-laboratory comparison of DNA fingerprinting data in the form of fluorescent

amplified fragment length polymorphism (FAFLP), enterobacterial repetitive

intergenic consensus (ERIC) and repetitive extragenic palindromic (REP) signature

profiles. Besides this, the database also displays diversity of strains based on

nucleotide sequences of several house keeping genes and two membrane proteins.

Being the first of its kind, genoBASE pylori is expected to be a helpful online

tool in strengthening the concept of 'geographic genomics' and will be useful to

molecular epidemiologists, clinical laboratory scientists and those interested in

diagnostic development for H. pylori. The database can be accessed through its

website (http://www.cdfd.org.in/amplibase/HP).

DOI: 10.1016/j.meegid.2007.01.006

PMID: 17320487 [Indexed for MEDLINE]

2632. J Comput Biol. 2007 Jul-Aug;14(6):839-55.

Simulating protein motions with rigidity analysis.

Thomas S(1), Tang X, Tapia L, Amato NM.

Author information:

(1)Parasol Lab, Department of Computer Science, Texas A&M University, College

Station, TX 77843-3112, USA.

Protein motions, ranging from molecular flexibility to large-scale conformational

change, play an essential role in many biochemical processes. Despite the

explosion in our knowledge of structural and functional data, our understanding

of protein movement is still very limited. In previous work, we developed and

validated a motion planning based method for mapping protein folding pathways

from unstructured conformations to the native state. In this paper, we propose a

novel method based on rigidity theory to sample conformation space more

effectively, and we describe extensions of our framework to automate the process

and to map transitions between specified conformations. Our results show that

these additions both improve the accuracy of our maps and enable us to study a

broader range of motions for larger proteins. For example, we show that

rigidity-based sampling results in maps that capture subtle folding differences

between protein G and its mutants, NuG1 and NuG2, and we illustrate how our

technique can be used to study large-scale conformational changes in calmodulin,

a 148 residue signaling protein known to undergo conformational changes when

binding to Ca(2+). Finally, we announce our web-based protein folding server

which includes a publicly available archive of protein motions:

(http://parasol.tamu.edu/foldingserver/).

DOI: 10.1089/cmb.2007.R019

PMID: 17691897 [Indexed for MEDLINE]

2633. J Mol Model. 2007 Jul;13(6-7):665-75. Epub 2007 Mar 30.

Localization of ligand binding site in proteins identified in silico.

Brylinski M(1), Kochanczyk M, Broniatowska E, Roterman I.

Author information:

(1)Department of Bioinformatics and Telemedicine, Jagiellonian

University-Collegium Medicum, Łazarza 16, 31-530, Krakow, Poland.

Knowledge-based models for protein folding assume that the early-stage structural

form of a polypeptide is determined by the backbone conformation, followed by

hydrophobic collapse. Side chain-side chain interactions, mostly of hydrophobic

character, lead to the formation of the hydrophobic core, which seems to

stabilize the structure of the protein in its natural environment. The

fuzzy-oil-drop model is employed to represent the idealized hydrophobicity

distribution in the protein molecule. Comparing it with the one empirically

observed in the protein molecule reveals that they are not in agreement. It is

shown in this study that the irregularity of hydrophobic distributions is

aim-oriented. The character and strength of these irregularities in the

organization of the hydrophobic core point to the specificity of a particular

protein's structure/function. When the location of these irregularities is

determined versus the idealized fuzzy-oil-drop, function-related areas in the

protein molecule can be identified. The presented model can also be used to

identify ways in which protein-protein complexes can possibly be created. Active

sites can be predicted for any protein structure according to the presented model

with the free prediction server at

http://www.bioinformatics.cm-uj.krakow.pl/activesite. The implication based on

the model presented in this work suggests the necessity of active presence of

ligand during the protein folding process simulation.

DOI: 10.1007/s00894-007-0191-x

PMID: 17394030 [Indexed for MEDLINE]

2634. Mol Biol (Mosk). 2007 Jul-Aug;41(4):711-8.

[A method for prediction of conserved RNA secondary structures].

[Article in Russian]

Mironov AA.

The RNA secondary structure prediction is a classical problem in bioinformatics.

The most efficient approach to this problem is based on the idea of a comparative

analysis. In this approach the algorithms utilize multiple alignment of the RNA

sequences and find common RNA structure. This paper describes a new algorithm for

this task. This algorithm does not require predefined multiple alignment. The

main idea of the algorithm is based on MEME-like iterative searching of abstract

profile on different levels. On the first level the algorithm searches the common

blocks in the RNA sequences and creates chain of this blocks. On the next step

the algorithm refines the chain of common blocks. On the last stage the algorithm

searches sets of common helices that have consistent locations relative to common

blocks. The algorithm was tested on sets of tRNA with a subset of junk sequences

and on RFN riboswitches. The algorithm is implemented as a web server

(http://bioinf.fbb.msu.ru/RNAAlign/).

PMID: 17936993 [Indexed for MEDLINE]

2635. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W58-62.

REPK: an analytical web server to select restriction endonucleases for terminal

restriction fragment length polymorphism analysis.

Collins RE(1), Rocap G.

Author information:

(1)School of Oceanography, University of Washington, Seattle, WA, USA.

Terminal restriction fragment length polymorphism (T-RFLP) analysis is a

widespread technique for rapidly fingerprinting microbial communities. Users of

T-RFLP frequently overlook the resolving power of well-chosen restriction

endonucleases and often fail to report how they chose their enzymes. REPK

(Restriction Endonuclease Picker) assists in the rational choice of restriction

endonucleases for T-RFLP by finding sets of four restriction endonucleases that

together uniquely differentiate user-designated sequence groups. With REPK, users

can provide their own sequences (of any gene, not just 16S rRNA), specify the

taxonomic rank of interest and choose from a number of filtering options to

further narrow down the enzyme selection. Bug tracking is provided, and the

source code is open and accessible under the GNU Public License v.2, at

http://code.google.com/p/repk. The web server is available without access

restrictions at http://rocaplab.ocean.washington.edu/tools/repk.

DOI: 10.1093/nar/gkm384

PMCID: PMC1933217

PMID: 17631616 [Indexed for MEDLINE]

2636. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W345-9.

CPC: assess the protein-coding potential of transcripts using sequence features

and support vector machine.

Kong L(1), Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G.

Author information:

(1)Center for Bioinformatics, National Laboratory of Protein Engineering and

Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing

100871, PR China.

Recent transcriptome studies have revealed that a large number of transcripts in

mammals and other organisms do not encode proteins but function as noncoding RNAs

(ncRNAs) instead. As millions of transcripts are generated by large-scale cDNA

and EST sequencing projects every year, there is a need for automatic methods to

distinguish protein-coding RNAs from noncoding RNAs accurately and quickly. We

developed a support vector machine-based classifier, named Coding Potential

Calculator (CPC), to assess the protein-coding potential of a transcript based on

six biologically meaningful sequence features. Tenfold cross-validation on the

training dataset and further testing on several large datasets showed that CPC

can discriminate coding from noncoding transcripts with high accuracy.

Furthermore, CPC also runs an order-of-magnitude faster than a previous

state-of-the-art tool and has higher accuracy. We developed a user-friendly

web-based interface of CPC at http://cpc.cbi.pku.edu.cn. In addition to

predicting the coding potential of the input transcripts, the CPC web server also

graphically displays detailed sequence features and additional annotations of the

transcript that may facilitate users' further investigation.

DOI: 10.1093/nar/gkm391

PMCID: PMC1933232

PMID: 17631615 [Indexed for MEDLINE]

2637. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W163-8. Epub 2007 Jul 10.

SAGExplore: a web server for unambiguous tag mapping in serial analysis of gene

expression oriented to gene discovery and annotation.

Norambuena T(1), Malig R, Melo F.

Author information:

(1)Departamento de Genética Molecular y Microbiología, Facultad de Ciencias

Biológicas, Pontificia Universidad Católica de Chile, Alameda 340, Santiago,

Chile.

We describe a web server for the accurate mapping of experimental tags in serial

analysis of gene expression (SAGE). The core of the server relies on a database

of genomic virtual tags built by a recently described method that attempts to

reduce the amount of ambiguous assignments for those tags that are not unique in

the genome. The method provides a complete annotation of potential virtual SAGE

tags within a genome, along with an estimation of their confidence for

experimental observation that ranks tags that present multiple matches in the

genome. The output of the server consists of a table in HTML format that contains

links to a graphic representation of the results and to some external servers and

databases, facilitating the tasks of analysis of gene expression and gene

discovery. Also, a table in tab delimited text format is produced, allowing the

user to export the results into custom databases and software for further

analysis. The current server version provides the most accurate and complete SAGE

tag mapping source that is available for the yeast organism. In the near future,

this server will also allow the accurate mapping of experimental SAGE-tags from

other model organisms such as human, mouse, frog and fly. The server is freely

available on the web at: http://dna.bio.puc.cl/SAGExplore.html.

DOI: 10.1093/nar/gkm429

PMCID: PMC1933165

PMID: 17626053 [Indexed for MEDLINE]

2638. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W350-3. Epub 2007 Jun 25.

TFAM 1.0: an online tRNA function classifier.

Tåquist H(1), Cui Y, Ardell DH.

Author information:

(1)Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, SE-751 24

Uppsala, Sweden.

We have earlier published an automated statistical classifier of tRNA function

called TFAM. Unlike tRNA gene-finders, TFAM uses information from the total

sequences of tRNAs and not just their anticodons to predict their function.

Therefore TFAM has an advantage in predicting initiator tRNAs, the amino acid

charging identity of nonstandard tRNAs such as suppressors, and the former

identity of pseudo-tRNAs. In addition, TFAM predictions are robust to sequencing

errors and useful for the statistical analysis of tRNA sequence, function and

evolution. Earlier versions of TFAM required a complicated installation and

running procedure, and only bacterial tRNA identity models were provided. Here we

describe a new version of TFAM with both a Web Server interface and simplified

standalone installation. New TFAM models are available including a

proteobacterial model for the bacterial lysylated isoleucine tRNAs, making it now

possible for TFAM to correctly classify all tRNA genes for some bacterial taxa.

First-draft eukaryotic and archaeal models are also provided making initiator

tRNA prediction easily accessible genes to any researcher or genome sequencing

effort. The TFAM Web Server is available at http://tfam.lcb.uu.se.

DOI: 10.1093/nar/gkm393

PMCID: PMC1933168

PMID: 17591612 [Indexed for MEDLINE]

2639. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W506-11. Epub 2007 Jun 22.

Selecton 2007: advanced models for detecting positive and purifying selection

using a Bayesian inference approach.

Stern A(1), Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T.

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, Ramat Aviv 69978, Israel.

Biologically significant sites in a protein may be identified by contrasting the

rates of synonymous (K(s)) and non-synonymous (K(a)) substitutions. This enables

the inference of site-specific positive Darwinian selection and purifying

selection. We present here Selecton version 2.2

(http://selecton.bioinfo.tau.ac.il), a web server which automatically calculates

the ratio between K(a) and K(s) (omega) at each site of the protein. This ratio

is graphically displayed on each site using a color-coding scheme, indicating

either positive selection, purifying selection or lack of selection. Selecton

implements an assembly of different evolutionary models, which allow for

statistical testing of the hypothesis that a protein has undergone positive

selection. Specifically, the recently developed mechanistic-empirical model is

introduced, which takes into account the physicochemical properties of amino

acids. Advanced options were introduced to allow maximal fine tuning of the

server to the user's specific needs, including calculation of statistical support

of the omega values, an advanced graphic display of the protein's 3-dimensional

structure, use of different genetic codes and inputting of a pre-built

phylogenetic tree. Selecton version 2.2 is an effective, user-friendly and freely

available web server which implements up-to-date methods for computing

site-specific selection forces, and the visualization of these forces on the

protein's sequence and structure.

DOI: 10.1093/nar/gkm382

PMCID: PMC1933148

PMID: 17586822 [Indexed for MEDLINE]

2640. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W3-5. Epub 2007 Jun 22.

Conducting research on the web: 2007 update for the bioinformatics links

directory.

Fox JA(1), McMillan S, Ouellette BF.

Author information:

(1)UBC Bioinformatics Centre, University of British Columbia, 2185 East Mall,

Vancouver, British Columbia, Canada, V6T 1Z4.

The Bioinformatics Links Directory, http://bioinformatics.ca/links\_directory, is

an actively maintained compilation of servers published in this and previous

issues of Nucleic Acids Research issues together with many other useful tools,

databases and resources for life sciences research. The 2007 update includes the

130 websites highlighted in the July 2007 Web Server issue of Nucleic Acids

Research and brings the total number of servers listed in the Bioinformatics

Links Directory to just under 1200 links. In addition to the updated content, the

2007 update of the Bioinformatics Links Directory includes new features for

improved navigation, accessibility and open data exchange. A complete listing of

all links listed in this Nucleic Acids Research 2007 Web Server issue can be

accessed online at, http://bioinformatics.ca/links\_directory/narweb2007. The 2007

update of the Bioinformatics Links Directory, which includes the Web Server list

and summaries is also available online, at the Nucleic Acids Research web site,

http://nar.oupjournals.org.

DOI: 10.1093/nar/gkm459

PMCID: PMC1933129

PMID: 17586821 [Indexed for MEDLINE]

2641. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W573-7. Epub 2007 Jun 21.

firestar--prediction of functionally important residues using structural

templates and alignment reliability.

López G(1), Valencia A, Tress ML.

Author information:

(1)Structural Biology and Biocomputing Program, Spanish National Cancer Research

Centre (CNIO) Melchor Fernández Almagro, 3, E-28029, Madrid, Spain.

glopez@cnio.es

Here we present firestar, an expert system for predicting ligand-binding residues

in protein structures. The server provides a method for extrapolating from the

large inventory of functionally important residues organized in the FireDB

database and adds information about the local conservation of potential-binding

residues. The interface allows users to make queries by protein sequence or

structure. The user can access pairwise and multiple alignments with structures

that have relevant functionally important binding sites. The results are

presented in a series of easy to read displays that allow users to compare

binding residue conservation across homologous proteins. The binding site

residues can also be viewed with molecular visualization tools. One feature of

firestar is that it can be used to evaluate the biological relevance of small

molecule ligands present in PDB structures. With the server it is easy to discern

whether small molecule binding is conserved in homologous structures. We found

this facility particularly useful during the recent assessment of CASP7 function

prediction.AVAILABILITY: http://firedb.bioinfo.cnio.es/Php/FireStar.php.

DOI: 10.1093/nar/gkm297

PMCID: PMC1933227

PMID: 17584799 [Indexed for MEDLINE]

2642. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W369-74. Epub 2007 Jun 21.

Pcons.net: protein structure prediction meta server.

Wallner B(1), Larsson P, Elofsson A.

Author information:

(1)Department of Biochemistry, University of Washington, Box 357350, Seattle, WA

98195, USA. bjornwa@u.washington.edu

The Pcons.net Meta Server (http://pcons.net) provides improved automated tools

for protein structure prediction and analysis using consensus. It essentially

implements all the steps necessary to produce a high quality model of a protein.

The whole process is fully automated and a potential user only submits the

protein sequence. For PSI-BLAST detectable targets, an accurate model is

generated within minutes of submission. For more difficult targets the sequence

is automatically submitted to publicly available fold-recognition servers that

use more advanced approaches to find distant structural homologs. The results

from these servers are analyzed and assessed for structural correctness using

Pcons and ProQ; and the user is presented with a ranked list of possible models.

In addition, if the protein sequence contains more than one domain, these are

automatically parsed out and resubmitted to the server as individual queries.

DOI: 10.1093/nar/gkm319

PMCID: PMC1933226

PMID: 17584798 [Indexed for MEDLINE]

2643. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W495-8. Epub 2007 Jun 21.

Sequence harmony: detecting functional specificity from alignments.

Feenstra KA(1), Pirovano W, Krab K, Heringa J.

Author information:

(1)Centre for Integrative Bioinformatics VU (IBIVU), Vrije Universiteit

Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands.

Multiple sequence alignments are often used for the identification of key

specificity-determining residues within protein families. We present a web server

implementation of the Sequence Harmony (SH) method previously introduced. SH

accurately detects subfamily specific positions from a multiple alignment by

scoring compositional differences between subfamilies, without imposing

conservation. The SH web server allows a quick selection of subtype specific

sites from a multiple alignment given a subfamily grouping. In addition, it

allows the predicted sites to be directly mapped onto a protein structure and

displayed. We demonstrate the use of the SH server using the family of plant

mitochondrial alternative oxidases (AOX). In addition, we illustrate the

usefulness of combining sequence and structural information by showing that the

predicted sites are clustered into a few distinct regions in an AOX homology

model. The SH web server can be accessed at www.ibi.vu.nl/programs/seqharmwww.

DOI: 10.1093/nar/gkm406

PMCID: PMC1933219

PMID: 17584793 [Indexed for MEDLINE]

2644. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W121-5. Epub 2007 Jun 21.

FGF: a web tool for Fishing Gene Family in a whole genome database.

Zheng H(1), Shi J, Fang X, Li Y, Vang S, Fan W, Wang J, Zhang Z, Wang W,

Kristiansen K, Wang J.

Author information:

(1)Beijing Institute of Genomics of Chinese Academy of Sciences, Beijing Genomics

Institute, Beijing 101300, China.

Gene duplication is an important process in evolution. The availability of genome

sequences of a number of organisms has made it possible to conduct comprehensive

searches for duplicated genes enabling informative studies of their evolution. We

have established the FGF (Fishing Gene Family) program to efficiently search for

and identify gene families. The FGF output displays the results as visual

phylogenetic trees including information on gene structure, chromosome position,

duplication fate and selective pressure. It is particularly useful to identify

pseudogenes and detect changes in gene structure. FGF is freely available on a

web server at http://fgf.genomics.org.cn/

DOI: 10.1093/nar/gkm426

PMCID: PMC1933194

PMID: 17584790 [Indexed for MEDLINE]

2645. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W619-24. Epub 2007 Jun 21.

DEEP--a tool for differential expression effector prediction.

Degenhardt J(1), Haubrock M, Dönitz J, Wingender E, Crass T.

Author information:

(1)Department of Bioinformatics, Medical Faculty, Georg August University,

Goldschmidtstrasse 1, 37077 Göttingen, Germany.

High-throughput methods for measuring transcript abundance, like SAGE or

microarrays, are widely used for determining differences in gene expression

between different tissue types, dignities (normal/malignant) or time points.

Further analysis of such data frequently aims at the identification of gene

interaction networks that form the causal basis for the observed properties of

the systems under examination. To this end, it is usually not sufficient to rely

on the measured gene expression levels alone; rather, additional biological

knowledge has to be taken into account in order to generate useful hypotheses

about the molecular mechanism leading to the realization of a certain phenotype.

We present a method that combines gene expression data with biological expert

knowledge on molecular interaction networks, as described by the TRANSPATH

database on signal transduction, to predict additional--and not necessarily

differentially expressed--genes or gene products which might participate in

processes specific for either of the examined tissues or conditions. In a first

step, significance values for over-expression in tissue/condition A or B are

assigned to all genes in the expression data set. Genes with a significance value

exceeding a certain threshold are used as starting points for the reconstruction

of a graph with signaling components as nodes and signaling events as edges. In a

subsequent graph traversal process, again starting from the previously identified

differentially expressed genes, all encountered nodes 'inherit' all their

starting nodes' significance values. In a final step, the graph is visualized,

the nodes being colored according to a weighted average of their inherited

significance values. Each node's, or sub-network's, predominant color, ranging

from green (significant for tissue/condition A) over yellow (not significant for

either tissue/condition) to red (significant for tissue/condition B), thus gives

an immediate visual clue on which molecules--differentially expressed or not--may

play pivotal roles in the tissues or conditions under examination. The described

method has been implemented in Java as a client/server application and a web

interface called DEEP (Differential Expression Effector Prediction). The client,

which features an easy-to-use graphical interface, can freely be downloaded from

the following URL: http://deep.bioinf.med.uni-goettingen.de.

DOI: 10.1093/nar/gkm469

PMCID: PMC1933247

PMID: 17584786 [Indexed for MEDLINE]

2646. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W455-9. Epub 2007 Jun 18.

The SLiMDisc server: short, linear motif discovery in proteins.

Davey NE(1), Edwards RJ, Shields DC.

Author information:

(1)UCD Conway Institute of Biomolecular and Biomedical Research, University

College Dublin, Belfield, Dublin 4, Ireland.

Short, linear motifs (SLiMs) play a critical role in many biological processes,

particularly in protein-protein interactions. Overrepresentation of convergent

occurrences of motifs in proteins with a common attribute (such as similar

subcellular location or a shared interaction partner) provides a feasible means

to discover novel occurrences computationally. The SLiMDisc (Short, Linear Motif

Discovery) web server corrects for common ancestry in describing shared motifs,

concentrating on the convergently evolved motifs. The server returns a listing of

the most interesting motifs found within unmasked regions, ranked according to an

information content-based scoring scheme. It allows interactive input masking,

according to various criteria. Scoring allows for evolutionary relationships in

the data sets through treatment of BLAST local alignments. Alongside this ranked

list, visualizations of the results improve understanding of the context of

suggested motifs, helping to identify true motifs of interest. These

visualizations include alignments of motif occurrences, alignments of motifs and

their homologues and a visual schematic of the top-ranked motifs. Additional

options for filtering and/or re-ranking motifs further permit the user to focus

on motifs with desired attributes. Returned motifs can also be compared with

known SLiMs from the literature. SLiMDisc is available at:

http://bioware.ucd.ie/~slimdisc/.

DOI: 10.1093/nar/gkm400

PMCID: PMC1933137

PMID: 17576682 [Indexed for MEDLINE]

2647. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W285-91. Epub 2007 Jun 18.

SplicePort--an interactive splice-site analysis tool.

Dogan RI(1), Getoor L, Wilbur WJ, Mount SM.

Author information:

(1)Computer Science Department, University of Maryland, College Park, Maryland

20742, USA. rezarta@cs.umd.edu

SplicePort is a web-based tool for splice-site analysis that allows the user to

make splice-site predictions for submitted sequences. In addition, the user can

also browse the rich catalog of features that underlies these predictions, and

which we have found capable of providing high classification accuracy on human

splice sites. Feature selection is optimized for human splice sites, but the

selected features are likely to be predictive for other mammals as well. With our

interactive feature browsing and visualization tool, the user can view and

explore subsets of features used in splice-site prediction (either the features

that account for the classification of a specific input sequence or the complete

collection of features). Selected feature sets can be searched, ranked or

displayed easily. The user can group features into clusters and frequency plot

WebLogos can be generated for each cluster. The user can browse the identified

clusters and their contributing elements, looking for new interesting signals, or

can validate previously observed signals. The SplicePort web server can be

accessed at http://www.cs.umd.edu/projects/SplicePort and

http://www.spliceport.org.

DOI: 10.1093/nar/gkm407

PMCID: PMC1933122

PMID: 17576680 [Indexed for MEDLINE]

2648. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W659-68. Epub 2007 Jun 13.

DIAL: a web server for the pairwise alignment of two RNA three-dimensional

structures using nucleotide, dihedral angle and base-pairing similarities.

Ferrè F(1), Ponty Y, Lorenz WA, Clote P.

Author information:

(1)Harvard Medical School, Children's Hospital, Hematology/Oncology Department,

Boston, MA 02115, USA.

DIAL (dihedral alignment) is a web server that provides public access to a new

dynamic programming algorithm for pairwise 3D structural alignment of RNA. DIAL

achieves quadratic time by performing an alignment that accounts for (i)

pseudo-dihedral and/or dihedral angle similarity, (ii) nucleotide sequence

similarity and (iii) nucleotide base-pairing similarity. DIAL provides access to

three alignment algorithms: global (Needleman-Wunsch), local (Smith-Waterman) and

semiglobal (modified to yield motif search). Suboptimal alignments are optionally

returned, and also Boltzmann pair probabilities Pr(a(i),b(j)) for aligned

positions a(i) , b(j) from the optimal alignment. If a non-zero suboptimal

alignment score ratio is entered, then the semiglobal alignment algorithm may be

used to detect structurally similar occurrences of a user-specified 3D motif. The

query motif may be contiguous in the linear chain or fragmented in a number of

noncontiguous regions. The DIAL web server provides graphical output which allows

the user to view, rotate and enlarge the 3D superposition for the optimal (and

suboptimal) alignment of query to target. Although graphical output is available

for all three algorithms, the semiglobal motif search may be of most interest in

attempts to identify RNA motifs. DIAL is available at

http://bioinformatics.bc.edu/clotelab/DIAL.

DOI: 10.1093/nar/gkm334

PMCID: PMC1933154

PMID: 17567620 [Indexed for MEDLINE]

2649. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W330-4. Epub 2007 Jun 12.

RNA Movies 2: sequential animation of RNA secondary structures.

Kaiser A(1), Krüger J, Evers DJ.

Author information:

(1)Faculty of Technology, Center for Biotechnology, Bielefeld University, 33594

Bielefeld, Germany.

RNA Movies is a simple, yet powerful visualization tool in likeness to a media

player application, which enables to browse sequential paths through RNA

secondary structure landscapes. It can be used to visualize structural

rearrangement processes of RNA, such as folding pathways and conformational

switches, or to browse lists of alternative structure candidates. Besides

extending the feature set, retaining and improving usability and availability in

the web is the main aim of this new version. RNA Movies now supports the DCSE and

RNAStructML input formats besides its own RNM format. Pseudoknots and 'entangled

helices' can be superimposed on the RNA secondary structure layout. Publication

quality output is provided through the Scalable Vector Graphics output format

understood by most current drawing programs. The software has been completely

re-implemented in Java to enable pure client-side operation as applet and

web-start application available at the Bielefeld Bioinformatics Server

http://bibiserv.techfak.uni-bielefeld.de/rnamovies.

DOI: 10.1093/nar/gkm309

PMCID: PMC1933240

PMID: 17567618 [Indexed for MEDLINE]

2650. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W398-402. Epub 2007 Jun 12.

eF-seek: prediction of the functional sites of proteins by searching for similar

electrostatic potential and molecular surface shape.

Kinoshita K(1), Murakami Y, Nakamura H.

Author information:

(1)Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai,

Minatoku, Tokyo, 108-8639, Japan. kino@ims.u-tokyo.ac.jp

We have developed a method to predict ligand-binding sites in a new protein

structure by searching for similar binding sites in the Protein Data Bank (PDB).

The similarities are measured according to the shapes of the molecular surfaces

and their electrostatic potentials. A new web server, eF-seek, provides an

interface to our search method. It simply requires a coordinate file in the PDB

format, and generates a prediction result as a virtual complex structure, with

the putative ligands in a PDB format file as the output. In addition, the

predicted interacting interface is displayed to facilitate the examination of the

virtual complex structure on our own applet viewer with the web browser (URL:

http://eF-site.hgc.jp/eF-seek).

DOI: 10.1093/nar/gkm351

PMCID: PMC1933152

PMID: 17567616 [Indexed for MEDLINE]

2651. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W314-9. Epub 2007 Jun 12.

RSRE: RNA structural robustness evaluator.

Shu W(1), Bo X, Zheng Z, Wang S.

Author information:

(1)Beijing Institute of Radiation Medicine, Beijing 100850, China.

Biological robustness, defined as the ability to maintain stable functioning in

the face of various perturbations, is an important and fundamental topic in

current biology, and has become a focus of numerous studies in recent years.

Although structural robustness has been explored in several types of RNA

molecules, the origins of robustness are still controversial. Computational

analysis results are needed to make up for the lack of evidence of robustness in

natural biological systems. The RNA structural robustness evaluator (RSRE) web

server presented here provides a freely available online tool to quantitatively

evaluate the structural robustness of RNA based on the widely accepted definition

of neutrality. Several classical structure comparison methods are employed; five

randomization methods are implemented to generate control sequences; sub-optimal

predicted structures can be optionally utilized to mitigate the uncertainty of

secondary structure prediction. With a user-friendly interface, the web

application is easy to use. Intuitive illustrations are provided along with the

original computational results to facilitate analysis. The RSRE will be helpful

in the wide exploration of RNA structural robustness and will catalyze our

understanding of RNA evolution. The RSRE web server is freely available at

http://biosrv1.bmi.ac.cn/RSRE/ or http://biotech.bmi.ac.cn/RSRE/.

DOI: 10.1093/nar/gkm361

PMCID: PMC1933138

PMID: 17567615 [Indexed for MEDLINE]

2652. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W460-4. Epub 2007 Jun 12.

PrDOS: prediction of disordered protein regions from amino acid sequence.

Ishida T(1), Kinoshita K.

Author information:

(1)Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai,

Minato-ku, Tokyo, 108-8639, Japan. t-ishida@hgc.jp

PrDOS is a server that predicts the disordered regions of a protein from its

amino acid sequence (http://prdos.hgc.jp). The server accepts a single protein

amino acid sequence, in either plain text or FASTA format. The prediction system

is composed of two predictors: a predictor based on local amino acid sequence

information and one based on template proteins. The server combines the results

of the two predictors and returns a two-state prediction (order/disorder) and a

disorder probability for each residue. The prediction results are sent by e-mail,

and the server also provides a web-interface to check the results.

DOI: 10.1093/nar/gkm363

PMCID: PMC1933209

PMID: 17567614 [Indexed for MEDLINE]

2653. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W465-72. Epub 2007 Jun 6.

iPDA: integrated protein disorder analyzer.

Su CT(1), Chen CY, Hsu CM.

Author information:

(1)Department of Computer Science and Information Engineering, National Taiwan

University, Taipei 106, Taiwan.

This article presents a web server iPDA, which aims at identifying the disordered

regions of a query protein. Automatic prediction of disordered regions from

protein sequences is an important problem in the study of structural biology. The

proposed classifier DisPSSMP2 is different from several existing disorder

predictors by its employment of position-specific scoring matrices with respect

to physicochemical properties (PSSMP), where the physicochemical properties

adopted here especially take the disorder propensity of amino acids into account.

The web server iPDA integrates DisPSSMP2 with several other sequence predictors

in order to investigate the functional role of the detected disordered region.

The predicted information includes sequence conservation, secondary structure,

sequence complexity and hydrophobic clusters. According to the proportion of the

secondary structure elements predicted, iPDA dynamically adjusts the cutting

threshold of determining protein disorder. Furthermore, a pattern mining package

for detecting sequence conservation is embedded in iPDA for discovering potential

binding regions of the query protein, which is really helpful to uncovering the

relationship between protein function and its primary sequence. The web service

is available at http://biominer.bime.ntu.edu.tw/ipda and mirrored at

http://biominer.cse.yzu.edu.tw/ipda.

DOI: 10.1093/nar/gkm353

PMCID: PMC1933224

PMID: 17553839 [Indexed for MEDLINE]

2654. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W339-44. Epub 2007 Jun 6.

MiPred: classification of real and pseudo microRNA precursors using random forest

prediction model with combined features.

Jiang P(1), Wu H, Wang W, Ma W, Sun X, Lu Z.

Author information:

(1)State Key Laboratory of Bioelectronics, Department of Biological Science and

Medical Engineering, Southeast University, Nanjing, 210096, PR China.

To distinguish the real pre-miRNAs from other hairpin sequences with similar

stem-loops (pseudo pre-miRNAs), a hybrid feature which consists of local

contiguous structure-sequence composition, minimum of free energy (MFE) of the

secondary structure and P-value of randomization test is used. Besides, a novel

machine-learning algorithm, random forest (RF), is introduced. The results

suggest that our method predicts at 98.21% specificity and 95.09% sensitivity.

When compared with the previous study, Triplet-SVM-classifier, our RF method was

nearly 10% greater in total accuracy. Further analysis indicated that the

improvement was due to both the combined features and the RF algorithm. The

MiPred web server is available at http://www.bioinf.seu.edu.cn/miRNA/. Given a

sequence, MiPred decides whether it is a pre-miRNA-like hairpin sequence or not.

If the sequence is a pre-miRNA-like hairpin, the RF classifier will predict

whether it is a real pre-miRNA or a pseudo one.

DOI: 10.1093/nar/gkm368

PMCID: PMC1933124

PMID: 17553836 [Indexed for MEDLINE]

2655. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W354-6. Epub 2007 Jun 6.

DOMAC: an accurate, hybrid protein domain prediction server.

Cheng J(1).

Author information:

(1)School of Electrical Engineering and Computer Science, University of Central

Florida, Orlando, FL 32816, USA. jcheng@cs.ucf.edu

Protein domain prediction is important for protein structure prediction,

structure determination, function annotation, mutagenesis analysis and protein

engineering. Here we describe an accurate protein domain prediction server

(DOMAC) combining both template-based and ab initio methods. The preliminary

version of the server was ranked among the top domain prediction servers in the

seventh edition of Critical Assessment of Techniques for Protein Structure

Prediction (CASP7), 2006. DOMAC server and datasets are available at:

http://www.bioinfotool.org/domac.html.

DOI: 10.1093/nar/gkm390

PMCID: PMC1933197

PMID: 17553833 [Indexed for MEDLINE]

2656. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W489-94. Epub 2007 Jun 6.

siteFiNDER|3D: a web-based tool for predicting the location of functional sites

in proteins.

Innis CA(1).

Author information:

(1)Howard Hughes Medical Institute/Yale University, Department of Molecular

Biophysics and Biochemistry, New Haven, CT 06520-8114, USA. axel.innis@yale.edu

Although knowledge of a protein's functional site is a key requirement for

understanding its mode of action at the molecular level, our ability to locate

such sites experimentally is far exceeded by the rate at which sequence and

structural information is being accumulated. siteFiNDER|3D is an online tool for

the prediction of functionally important regions in proteins of known structure.

At the core of the server lies the CFG analysis algorithm, which uses a moving 3D

window to correlate patterns of functional/chemical group conservation in the

query protein with the location of functional sites. Here, we give a general

overview of the functionality offered by the siteFiNDER|3D server, along with

general recommendations aimed at maximizing the accuracy and predictive value of

this tool in a variety of contexts. siteFiNDER|3D can be accessed at:

'http://sage.csb.yale.edu/sitefinder3d' and requires, at a minimum, the atomic

coordinates of a query protein in PDB format.

DOI: 10.1093/nar/gkm422

PMCID: PMC1933183

PMID: 17553829 [Indexed for MEDLINE]

2657. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W707-12. Epub 2007 Jun 6.

The Multi-Q web server for multiplexed protein quantitation.

Yu CY(1), Tsui YH, Yian YH, Sung TY, Hsu WL.

Author information:

(1)Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC.

The Multi-Q web server provides an automated data analysis tool for multiplexed

protein quantitation based on the iTRAQ labeling method. The web server is

designed as a platform that can accommodate various input data formats from

search engines and mass spectrometer manufacturers. Compared to the previous

stand-alone version, the new web server version provides many enhanced features

and flexible options for quantitation. The workflow of the web server is

represented by a quantitation wizard so that the tool is easy to use. It also

provides a friendly interface that helps users configure their parameter settings

before running the program. The web server generates a standard report for

quantitation results. In addition, it allows users to customize their output

reports and information of interest can be easily highlighted. The output also

provides visualization of mass spectral data so that users can conveniently

validate the results. The Multi-Q web server is a fully automated and easy to use

quantitation tool that is suitable for large-scale multiplexed protein

quantitation. Users can download the Multi-Q Web Server from

http://ms.iis.sinica.edu.tw/Multi-Q-Web.

DOI: 10.1093/nar/gkm345

PMCID: PMC1933177

PMID: 17553828 [Indexed for MEDLINE]

2658. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W433-7. Epub 2007 Jun 1.

MyHits: improvements to an interactive resource for analyzing protein sequences.

Pagni M(1), Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Hau J, Martin O,

Kuznetsov D, Falquet L.

Author information:

(1)Swiss Institute of Bioinformatics (SIB), Vital-IT Group, UNIL-Génopode,

CH-1015 Lausanne, Switzerland.

The MyHits web site (http://myhits.isb-sib.ch) is an integrated service dedicated

to the analysis of protein sequences. Since its first description in 2004, both

the user interface and the back end of the server were improved. A number of

tools (e.g. MAFFT, Jacop, Dotlet, Jalview, ESTScan) were added or updated to

improve the usability of the service. The MySQL schema and its associated API

were revamped and the database engine (HitKeeper) was separated from the web

interface. This paper summarizes the current status of the server, with an

emphasis on the new services.

DOI: 10.1093/nar/gkm352

PMCID: PMC1933190

PMID: 17545200 [Indexed for MEDLINE]

2659. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W280-4. Epub 2007 Jun 1.

FLAN: a web server for influenza virus genome annotation.

Bao Y(1), Bolotov P, Dernovoy D, Kiryutin B, Tatusova T.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, Bethesda, MD 20894, USA.

FLAN (short for FLu ANnotation), the NCBI web server for genome annotation of

influenza virus (http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/annotation.cgi)

is a tool for user-provided influenza A virus or influenza B virus sequences. It

can validate and predict protein sequences encoded by an input flu sequence. The

input sequence is BLASTed against a database containing influenza sequences to

determine the virus type (A or B), segment (1 through 8) and subtype for the

hemagglutinin and neuraminidase segments of influenza A virus. For each

segment/subtype of the viruses, a set of sample protein sequences is maintained.

The input sequence is then aligned against the corresponding protein set with a

'Protein to nucleotide alignment tool' (ProSplign). The translated product from

the best alignment to the sample protein sequence is used as the predicted

protein encoded by the input sequence. The output can be a feature table that can

be used for sequence submission to GenBank (by Sequin or tbl2asn), a GenBank flat

file, or the predicted protein sequences in FASTA format. A message showing the

length of the input sequence, the predicted virus type, segment and subtype for

the hemagglutinin and neuraminidase segments of Influenza A virus will also be

displayed.

DOI: 10.1093/nar/gkm354

PMCID: PMC1933127

PMID: 17545199 [Indexed for MEDLINE]

2660. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W105-14. Epub 2007 Jun 1.

Cross-species microarray analysis with the OSCAR system suggests an

INSR->Pax6->NQO1 neuro-protective pathway in aging and Alzheimer's disease.

Lu Y(1), He X, Zhong S.

Author information:

(1)Department of Computer Science, University of Illinois at Urbana-Champaign,

IL, USA.

OSCAR is a web platform for cluster and cross-species analysis of microarray

data. It provides a comprehensive but friendly environment to both users and

algorithm developers. For users, OSCAR provides cluster tools for both single and

multiple species data, together with interactive analysis features. For single

species data, OSCAR currently provides Hierarchical Clustering, K-means,

partition around medoids (PAM), Self-Organizing Map (SOM), Tight Clustering and a

novel algorithm called 'Consensus Tight-clustering'. The new Consensus

Tight-clustering algorithm delivers robust gene clusters and its result is more

resistant to false positives than other state-of-the-art algorithms. For

cross-species data analysis, OSCAR provides two novel computational tools:

'coherentCluster', 'coherentSubset' and a novel visualization tool: 'comparative

heatmap'. Applying the coherentCluster algorithm to human and fly aging data, we

identified several coherent clusters of genes, which share co-regulation patterns

that are highly correlated with the aging process in both of the two species. One

coherent cluster suggests insulin receptor (INSR) may regulate Pax6 in both

species and across different tissues. Further analysis with human brain

expression and pathological data suggests an INSR->Pax6->quinone oxidoreductase

(NQO1)->detoxification neuro-protective pathway might be present in aging or

diseased brain. For algorithm developers, OSCAR is a plug-and-play platform. With

little effort, developers can plug their own algorithms into the OSCAR server

without revealing the source codes, which will equip their command line

executables with user-friendly interface and interactive analysis capability. In

summary, OSCAR initiates an open platform for development and application of

clustering and cross-species analysis programs. OSCAR stands for an open system

for cluster analysis of microarray data. It is available at:

http://biocomp.bioen.uiuc.edu/oscar.

DOI: 10.1093/nar/gkm408

PMCID: PMC1933158

PMID: 17545194 [Indexed for MEDLINE]

2661. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W384-92. Epub 2007 May 30.

Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein

structures with linkage to metabolic pathways.

Uzun A(1), Leslin CM, Abyzov A, Ilyin V.

Author information:

(1)Department of Biology, Northeastern University, 360 Huntington Ave., Boston,

MA 02115, USA.

SNPs located within the open reading frame of a gene that result in an alteration

in the amino acid sequence of the encoded protein [nonsynonymous SNPs (nsSNPs)]

might directly or indirectly affect functionality of the protein, alone or in the

interactions in a multi-protein complex, by increasing/decreasing the activity of

the metabolic pathway. Understanding the functional consequences of such changes

and drawing conclusions about the molecular basis of diseases, involves

integrating information from multiple heterogeneous sources including sequence,

structure data and pathway relations between proteins. The data from NCBI's SNP

database (dbSNP), gene and protein databases from Entrez, protein structures from

the PDB and pathway information from KEGG have all been cross referenced into the

StSNP web server, in an effort to provide combined integrated, reports about

nsSNPs. StSNP provides 'on the fly' comparative modeling of nsSNPs with links to

metabolic pathway information, along with real-time visual comparative analysis

of the modeled structures using the Friend software application. The use of

metabolic pathways in StSNP allows a researcher to examine possible

disease-related pathways associated with a particular nsSNP(s), and link the

diseases with the current available molecular structure data. The server is

publicly available at http://glinka.bio.neu.edu/StSNP/.

DOI: 10.1093/nar/gkm232

PMCID: PMC1933130

PMID: 17537826 [Indexed for MEDLINE]

2662. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W688-93. Epub 2007 May 30.

BioBayesNet: a web server for feature extraction and Bayesian network modeling of

biological sequence data.

Nikolajewa S(1), Pudimat R, Hiller M, Platzer M, Backofen R.

Author information:

(1)Department of Bioinformatics, Friedrich-Schiller-University Jena,

Ernst-Abbe-Platz 2, 07743 Jena, Germany.

BioBayesNet is a new web application that allows the easy modeling and

classification of biological data using Bayesian networks. To learn Bayesian

networks the user can either upload a set of annotated FASTA sequences or a set

of pre-computed feature vectors. In case of FASTA sequences, the server is able

to generate a wide range of sequence and structural features from the sequences.

These features are used to learn Bayesian networks. An automatic feature

selection procedure assists in selecting discriminative features, providing an

(locally) optimal set of features. The output includes several quality measures

of the overall network and individual features as well as a graphical

representation of the network structure, which allows to explore dependencies

between features. Finally, the learned Bayesian network or another uploaded

network can be used to classify new data. BioBayesNet facilitates the use of

Bayesian networks in biological sequences analysis and is flexible to support

modeling and classification applications in various scientific fields. The

BioBayesNet server is available at

http://biwww3.informatik.uni-freiburg.de:8080/BioBayesNet/.

DOI: 10.1093/nar/gkm292

PMCID: PMC1933181

PMID: 17537825 [Indexed for MEDLINE]

2663. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W52-7. Epub 2007 May 30.

CRISPRFinder: a web tool to identify clustered regularly interspaced short

palindromic repeats.

Grissa I(1), Vergnaud G, Pourcel C.

Author information:

(1)Univ Paris-Sud, Institut de Génétique et Microbiologie, UMR 8621, Orsay,

F-91405, France. Ibtissem.Grissa@igmors.u-psud.fr

Clustered regularly interspaced short palindromic repeats (CRISPRs) constitute a

particular family of tandem repeats found in a wide range of prokaryotic genomes

(half of eubacteria and almost all archaea). They consist of a succession of

highly conserved regions (DR) varying in size from 23 to 47 bp, separated by

similarly sized unique sequences (spacer) of usually viral origin. A CRISPR

cluster is flanked on one side by an AT-rich sequence called the leader and

assumed to be a transcriptional promoter. Recent studies suggest that this

structure represents a putative RNA-interference-based immune system. Here we

describe CRISPRFinder, a web service offering tools to (i) detect CRISPRs

including the shortest ones (one or two motifs); (ii) define DRs and extract

spacers; (iii) get the flanking sequences to determine the leader; (iv) blast

spacers against Genbank database and (v) check if the DR is found elsewhere in

prokaryotic sequenced genomes. CRISPRFinder is freely accessible at

http://crispr.u-psud.fr/Server/CRISPRfinder.php.

DOI: 10.1093/nar/gkm360

PMCID: PMC1933234

PMID: 17537822 [Indexed for MEDLINE]

2664. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W227-31. Epub 2007 May 30.

Melina II: a web tool for comparisons among several predictive algorithms to find

potential motifs from promoter regions.

Okumura T(1), Makiguchi H, Makita Y, Yamashita R, Nakai K.

Author information:

(1)Mitsui Knowledge Industry Co. Ltd, University of Tokyo, Japan.

We present the second version of Melina, a web-based tool for promoter analysis.

Melina II shows potential DNA motifs in promoter regions with a combination of

several available programs, Consensus, MEME, Gibbs sampler, MDscan and Weeder, as

well as several parameter settings. It allows running a maximum of four programs

simultaneously, and comparing their results with graphical representations. In

addition, users can build a weight matrix from a predicted motif and apply it to

upstream sequences of several typical genomes (human, mouse, S. cerevisiae, E.

coli, B. subtilis or A. thaliana) or to public motif databases (JASPAR or DBTBS)

in order to find similar motifs. Melina II is a client/server system developed by

using Adobe (Macromedia) Flash and is accessible over the web at

http://melina.hgc.jp.

DOI: 10.1093/nar/gkm362

PMCID: PMC1933176

PMID: 17537821 [Indexed for MEDLINE]

2665. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W275-9. Epub 2007 May 30.

FluGenome: a web tool for genotyping influenza A virus.

Lu G(1), Rowley T, Garten R, Donis RO.

Author information:

(1)Department of Biology, University of Nebraska at Omaha, Omaha, NE, USA.

Influenza A viruses are hosted by numerous avian and mammalian species, which

have shaped their evolution into distinct lineages worldwide. The viral genome

consists of eight RNA segments that are frequently exchanged between different

viruses via a process known as genetic reassortment. A complete genotype

nomenclature is essential to describe gene segment reassortment. Specialized

bioinformatic tools to analyze reassortment are not available, which hampers

progress in understanding its role in host range, virulence and transmissibility

of influenza viruses. To meet this need, we have developed a nomenclature to name

influenza A genotypes and implemented a web server, FluGenome

(http://www.flugenome.org/), for the assignment of lineages and genotypes.

FluGenome provides functions for the user to interrogate the database in

different modalities and get detailed reports on lineages and genotypes. These

features make FluGenome unique in its ability to automatically detect genotype

differences attributable to reassortment events in influenza A virus evolution.

DOI: 10.1093/nar/gkm365

PMCID: PMC1933150

PMID: 17537820 [Indexed for MEDLINE]

2666. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W694-700. Epub 2007 May 30.

EVALLER: a web server for in silico assessment of potential protein

allergenicity.

Martinez Barrio A(1), Soeria-Atmadja D, Nistér A, Gustafsson MG, Hammerling U,

Bongcam-Rudloff E.

Author information:

(1)Linnaeus Centre for Bioinformatics, Uppsala Biomedical Centre (BMC), Uppsala

University, P.O. Box 598, SE-751 24 Uppsala, Sweden.

Bioinformatics testing approaches for protein allergenicity, involving amino acid

sequence comparisons, have evolved appreciably over the last several years to

increased sophistication and performance. EVALLER, the web server presented in

this article is based on our recently published 'Detection based on Filtered

Length-adjusted Allergen Peptides' (DFLAP) algorithm, which affords in silico

determination of potential protein allergenicity of high sensitivity and

excellent specificity. To strengthen bioinformatics risk assessment in

allergology EVALLER provides a comprehensive outline of its judgment on a query

protein's potential allergenicity. Each such textual output incorporates a

scoring figure, a confidence numeral of the assignment and information on high-

or low-scoring matches to identified allergen-related motifs, including their

respective location in accordingly derived allergens. The interface, built on a

modified Perl Open Source package, enables dynamic and color-coded graphic

representation of key parts of the output. Moreover, pertinent details can be

examined in great detail through zoomed views. The server can be accessed at

http://bioinformatics.bmc.uu.se/evaller.html.

DOI: 10.1093/nar/gkm370

PMCID: PMC1933222

PMID: 17537818 [Indexed for MEDLINE]

2667. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W221-6. Epub 2007 May 30.

MYBS: a comprehensive web server for mining transcription factor binding sites in

yeast.

Tsai HK(1), Chou MY, Shih CH, Huang GT, Chang TH, Li WH.

Author information:

(1)Institute of Information Science, Academia Sinica, Taipei, 115 Taiwan.

hktsai@iis.sinica.edu.tw

Correct interactions between transcription factors (TFs) and their binding sites

(TFBSs) are of central importance to gene regulation. Recently developed

chromatin-immunoprecipitation DNA chip (ChIP-chip) techniques and the

phylogenetic footprinting method provide ways to identify TFBSs with high

precision. In this study, we constructed a user-friendly interactive platform for

dynamic binding site mapping using ChIP-chip data and phylogenetic footprinting

as two filters. MYBS (Mining Yeast Binding Sites) is a comprehensive web server

that integrates an array of both experimentally verified and predicted position

weight matrixes (PWMs) from eleven databases, including 481 binding motif

consensus sequences and 71 PWMs that correspond to 183 TFs. MYBS users can search

within this platform for motif occurrences (possible binding sites) in the

promoters of genes of interest via simple motif or gene queries in conjunction

with the above two filters. In addition, MYBS enables users to visualize in

parallel the potential regulators for a given set of genes, a feature useful for

finding potential regulatory associations between TFs. MYBS also allows users to

identify target gene sets of each TF pair, which could be used as a starting

point for further explorations of TF combinatorial regulation. MYBS is available

at http://cg1.iis.sinica.edu.tw/~mybs/.

DOI: 10.1093/nar/gkm379

PMCID: PMC1933147

PMID: 17537814 [Indexed for MEDLINE]

2668. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W97-W104. Epub 2007 May 30.

MobilomeFINDER: web-based tools for in silico and experimental discovery of

bacterial genomic islands.

Ou HY(1), He X, Harrison EM, Kulasekara BR, Thani AB, Kadioglu A, Lory S, Hinton

JC, Barer MR, Deng Z, Rajakumar K.

Author information:

(1)Laboratory of Microbial Metabolism and School of Life Science & Biotechnology,

Shanghai Jiaotong University, PR China.

MobilomeFINDER (http://mml.sjtu.edu.cn/MobilomeFINDER) is an interactive online

tool that facilitates bacterial genomic island or 'mobile genome' (mobilome)

discovery; it integrates the ArrayOme and tRNAcc software packages. ArrayOme

utilizes a microarray-derived comparative genomic hybridization input data set to

generate 'inferred contigs' produced by merging adjacent genes classified as

'present'. Collectively these 'fragments' represent a hypothetical

'microarray-visualized genome (MVG)'. ArrayOme permits recognition of

discordances between physical genome and MVG sizes, thereby enabling

identification of strains rich in microarray-elusive novel genes. Individual

tRNAcc tools facilitate automated identification of genomic islands by

comparative analysis of the contents and contexts of tRNA sites and other

integration hotspots in closely related sequenced genomes. Accessory tools

facilitate design of hotspot-flanking primers for in silico and/or

wet-science-based interrogation of cognate loci in unsequenced strains and

analysis of islands for features suggestive of foreign origins; island-specific

and genome-contextual features are tabulated and represented in schematic and

graphical forms. To date we have used MobilomeFINDER to analyse several

Enterobacteriaceae, Pseudomonas aeruginosa and Streptococcus suis genomes.

MobilomeFINDER enables high-throughput island identification and characterization

through increased exploitation of emerging sequence data and PCR-based profiling

of unsequenced test strains; subsequent targeted yeast recombination-based

capture permits full-length sequencing and detailed functional studies of novel

genomic islands.

DOI: 10.1093/nar/gkm380

PMCID: PMC1933208

PMID: 17537813 [Indexed for MEDLINE]

2669. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W526-30. Epub 2007 May 30.

Patch Finder Plus (PFplus): a web server for extracting and displaying positive

electrostatic patches on protein surfaces.

Shazman S(1), Celniker G, Haber O, Glaser F, Mandel-Gutfreund Y.

Author information:

(1)Department of Biology, Technion-Israel Institute of Technology, Haifa 32000,

Israel.

Positively charged electrostatic patches on protein surfaces are usually

indicative of nucleic acid binding interfaces. Interestingly, many proteins which

are not involved in nucleic acid binding possess large positive patches on their

surface as well. In some cases, the positive patches on the protein are related

to other functional properties of the protein family. PatchFinderPlus (PFplus)

http://pfp.technion.ac.il is a web-based tool for extracting and displaying

continuous electrostatic positive patches on protein surfaces. The input required

for PFplus is either a four letter PDB code or a protein coordinate file in PDB

format, provided by the user. PFplus computes the continuum electrostatics

potential and extracts the largest positive patch for each protein chain in the

PDB file. The server provides an output file in PDB format including a list of

the patch residues. In addition, the largest positive patch is displayed on the

server by a graphical viewer (Jmol), using a simple color coding.

DOI: 10.1093/nar/gkm401

PMCID: PMC1933175

PMID: 17537808 [Indexed for MEDLINE]

2670. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W538-42. Epub 2007 May 25.

CytoSVM: an advanced server for identification of cytokine-receptor interactions.

Xu JR(1), Zhang JX, Han BC, Liang L, Ji ZL.

Author information:

(1)Key Laboratory for Cell Biology & Tumor Cell Engineering, the Ministry of

Education of China, School of Life Sciences, Xiamen University, Xiamen 361005,

FuJian Province, PR China.

The interactions between cytokines and their complementary receptors are the

gateways to properly understand a large variety of cytokine-specific cellular

activities such as immunological responses and cell differentiation. To discover

novel cytokine-receptor interactions, an advanced support vector machines (SVMs)

model, CytoSVM, was constructed in this study. This model was iteratively trained

using 449 mammal (except rat) cytokine-receptor interactions and about 1 million

virtually generated positive and negative vectors in an enriched way. Final

independent evaluation by rat's data received sensitivity of 97.4%, specificity

of 99.2% and the Matthews correlation coefficient (MCC) of 0.89. This performance

is better than normal SVM-based models. Upon this well-optimized model, a

web-based server was created to accept primary protein sequence and present its

probabilities to interact with one or several cytokines. Moreover, this model was

applied to identify putative cytokine-receptor pairs in the whole genomes of

human and mouse. Excluding currently known cytokine-receptor interactions, total

1609 novel cytokine-receptor pairs were discovered from human genome with

probability approximately 80% after further transmembrane analysis. These cover

220 novel receptors (excluding their isoforms) for 126 human cytokines. The

screening results have been deposited in a database. Both the server and the

database can be freely accessed at

http://bioinf.xmu.edu.cn/software/cytosvm/cytosvm.php.

DOI: 10.1093/nar/gkm254

PMCID: PMC1933174

PMID: 17526528 [Indexed for MEDLINE]

2671. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W305-9. Epub 2007 May 25.

RNAbor: a web server for RNA structural neighbors.

Freyhult E(1), Moulton V, Clote P.

Author information:

(1)Linnaeus Centre for Bioinformatics, Uppsala University, 75124 Uppsala, Sweden.

RNAbor provides a new tool for researchers in the biological and related sciences

to explore important aspects of RNA secondary structure and folding pathways.

RNAbor computes statistics concerning delta-neighbors of a given input RNA

sequence and structure (the structure can, for example, be the minimum free

energy (MFE) structure). A delta-neighbor is a structure that differs from the

input structure by exactly delta base pairs, that is, it can be obtained from the

input structure by adding and/or removing exactly delta base pairs. For each

distance delta RNAbor computes the density of delta-neighbors, the number of

delta-neighbors, and the MFE structure, or MFE (delta) structure, among all

delta-neighbors. RNAbor can be used to study possible folding pathways, to

determine alternate low-energy structures, to predict potential nucleation sites

and to explore structural neighbors of an intermediate, biologically active

structure. The web server is available at

http://bioinformatics.bc.edu/clotelab/RNAbor.

DOI: 10.1093/nar/gkm255

PMCID: PMC1933207

PMID: 17526527 [Indexed for MEDLINE]

2672. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W512-21. Epub 2007 May 25.

A server and database for dipole moments of proteins.

Felder CE(1), Prilusky J, Silman I, Sussman JL.

Author information:

(1)Department of Structural Biology, Weizmann Institute of Science, 76100

Rehovot, Israel.

An Internet server at http://bip.weizmann.ac.il/dipol calculates the net charge,

dipole moment and mean radius of any 3D protein structure or its constituent

peptide chains, and displays the dipole vector superimposed on a ribbon backbone

of the protein. The server can also display the angle between the dipole and a

selected list of amino acid residues in the protein. When the net charges and

dipole moments of approximately 12 000 non-homologous PDB biological units

(PISCES set), and their unique chains of length 50 residues or longer, were

examined, the great majority of both charges and dipoles fell into a very narrow

range of values, with long extended tails containing a few extreme outliers. In

general, there is no obvious relation between a protein's charge or dipole moment

and its structure or function, so that its electrostatic properties are highly

specific to the particular protein, except that the majority of chains with very

large positive charges or dipoles bind to ribosomes or interact with nucleic

acids.

DOI: 10.1093/nar/gkm307

PMCID: PMC1933167

PMID: 17526523 [Indexed for MEDLINE]

2673. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W182-5. Epub 2007 May 25.

KAAS: an automatic genome annotation and pathway reconstruction server.

Moriya Y(1), Itoh M, Okuda S, Yoshizawa AC, Kanehisa M.

Author information:

(1)Bioinformatics Center, Institute for Chemical Research, Kyoto University,

Gokasho, Uji, Kyoto 611-0011, Japan.

The number of complete and draft genomes is rapidly growing in recent years, and

it has become increasingly important to automate the identification of functional

properties and biological roles of genes in these genomes. In the KEGG database,

genes in complete genomes are annotated with the KEGG orthology (KO) identifiers,

or the K numbers, based on the best hit information using Smith-Waterman scores

as well as by the manual curation. Each K number represents an ortholog group of

genes, and it is directly linked to an object in the KEGG pathway map or the

BRITE functional hierarchy. Here, we have developed a web-based server called

KAAS (KEGG Automatic Annotation Server: http://www.genome.jp/kegg/kaas/) i.e. an

implementation of a rapid method to automatically assign K numbers to genes in

the genome, enabling reconstruction of KEGG pathways and BRITE hierarchies. The

method is based on sequence similarities, bi-directional best hit information and

some heuristics, and has achieved a high degree of accuracy when compared with

the manually curated KEGG GENES database.

DOI: 10.1093/nar/gkm321

PMCID: PMC1933193

PMID: 17526522 [Indexed for MEDLINE]

2674. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W186-92. Epub 2007 May 25.

GeneTrail--advanced gene set enrichment analysis.

Backes C(1), Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Müller R,

Meese E, Lenhof HP.

Author information:

(1)Center for Bioinformatics, Saarland University, Building E1.1, 66041

Saarbrücken, Germany. cbackes@bioinf.uni-sb.de

We present a comprehensive and efficient gene set analysis tool, called

'GeneTrail' that offers a rich functionality and is easy to use. Our web-based

application facilitates the statistical evaluation of high-throughput genomic or

proteomic data sets with respect to enrichment of functional categories.

GeneTrail covers a wide variety of biological categories and pathways, among

others KEGG, TRANSPATH, TRANSFAC, and GO. Our web server provides two common

statistical approaches, 'Over-Representation Analysis' (ORA) comparing a

reference set of genes to a test set, and 'Gene Set Enrichment Analysis' (GSEA)

scoring sorted lists of genes. Besides other newly developed features,

GeneTrail's statistics module includes a novel dynamic-programming algorithm that

improves the P-value computation of GSEA methods considerably. GeneTrail is

freely accessible at http://genetrail.bioinf.uni-sb.de.

DOI: 10.1093/nar/gkm323

PMCID: PMC1933132

PMID: 17526521 [Indexed for MEDLINE]

2675. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W645-8. Epub 2007 May 25.

The M-Coffee web server: a meta-method for computing multiple sequence alignments

by combining alternative alignment methods.

Moretti S(1), Armougom F, Wallace IM, Higgins DG, Jongeneel CV, Notredame C.

Author information:

(1)Swiss Institute of Bioinformatics, Bâtiment Génopode, UNIL, CH-101 Lausanne.

The M-Coffee server is a web server that makes it possible to compute multiple

sequence alignments (MSAs) by running several MSA methods and combining their

output into one single model. This allows the user to simultaneously run all his

methods of choice without having to arbitrarily choose one of them. The MSA is

delivered along with a local estimation of its consistency with the individual

MSAs it was derived from. The computation of the consensus multiple alignment is

carried out using a special mode of the T-Coffee package [Notredame, Higgins and

Heringa (T-Coffee: a novel method for fast and accurate multiple sequence

alignment. J. Mol. Biol. 2000; 302: 205-217); Wallace, O'Sullivan, Higgins and

Notredame (M-Coffee: combining multiple sequence alignment methods with T-Coffee.

Nucleic Acids Res. 2006; 34: 1692-1699)] Given a set of sequences (DNA or

proteins) in FASTA format, M-Coffee delivers a multiple alignment in the most

common formats. M-Coffee is a freeware open source package distributed under a

GPL license and it is available either as a standalone package or as a web

service from www.tcoffee.org.

DOI: 10.1093/nar/gkm333

PMCID: PMC1933118

PMID: 17526519 [Indexed for MEDLINE]

2676. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W201-5. Epub 2007 May 25.

MADAP, a flexible clustering tool for the interpretation of one-dimensional

genome annotation data.

Schmid CD(1), Sengstag T, Bucher P, Delorenzi M.

Author information:

(1)Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland.

christoph.schmid@isb-sib.ch

A recurring task in the analysis of mass genome annotation data from

high-throughput technologies is the identification of peaks or clusters in a

noisy signal profile. Examples of such applications are the definition of

promoters on the basis of transcription start site profiles, the mapping of

transcription factor binding sites based on ChIP-chip data and the identification

of quantitative trait loci (QTL) from whole genome SNP profiles. Input to such an

analysis is a set of genome coordinates associated with counts or intensities.

The output consists of a discrete number of peaks with respective volumes,

extensions and center positions. We have developed for this purpose a flexible

one-dimensional clustering tool, called MADAP, which we make available as a web

server and as standalone program. A set of parameters enables the user to

customize the procedure to a specific problem. The web server, which returns

results in textual and graphical form, is useful for small to medium-scale

applications, as well as for evaluation and parameter tuning in view of

large-scale applications, requiring a local installation. The program written in

C++ can be freely downloaded from ftp://ftp.epd.unil.ch/pub/software/unix/madap.

The MADAP web server can be accessed at http://www.isrec.isb-sib.ch/madap/.

DOI: 10.1093/nar/gkm343

PMCID: PMC1933235

PMID: 17526516 [Indexed for MEDLINE]

2677. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W444-50. Epub 2007 May 25.

ProtSweep, 2Dsweep and DomainSweep: protein analysis suite at DKFZ.

del Val C(1), Ernst P, Falkenhahn M, Fladerer C, Glatting KH, Suhai S,

Hotz-Wagenblatt A.

Author information:

(1)DKFZ, German Cancer Research Center, Division of Molecular Biophysics, Im

Neuenheimer Feld 580, D-69120 Heidelberg, Germany. c.delval@dkfz.de

The wealth of transcript information that has been made publicly available in

recent years has led to large pools of individual web sites offering access to

bioinformatics software. However, finding out which services exist, what they can

or cannot do, how to use them and how to feed results from one service to the

next one in the right format can be very time and resource consuming, especially

for non-experts. Automating this task, we present a suite of protein annotation

pipelines (tasks) developed at the German Cancer Research Centre (DKFZ) oriented

to protein annotation by homology (ProtSweep), by domain analysis (DomainSweep),

and by secondary structure elements (2Dsweep). The aim of these tasks is to

perform an exhaustive structural and functional analysis employing a wide variety

of methods in combination with the most updated public databases. The three

servers are available for academic users at the HUSAR open server

http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar/

DOI: 10.1093/nar/gkm364

PMCID: PMC1933246

PMID: 17526514 [Indexed for MEDLINE]

2678. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W159-62. Epub 2007 May 25.

ESTpass: a web-based server for processing and annotating expressed sequence tag

(EST) sequences.

Lee B(1), Hong T, Byun SJ, Woo T, Choi YJ.

Author information:

(1)Korean BioInformation Center, KRIBB, Daejeon 305-817, Korea.

We present a web-based server, called ESTpass, for processing and annotating

sequence data from expressed sequence tag (EST) projects. ESTpass accepts a

FASTA-formatted EST file and its quality file as inputs, and it then executes a

back-end EST analysis pipeline consisting of three consecutive steps. The first

is cleansing the input EST sequences. The second is clustering and assembling the

cleansed EST sequences using d2\_cluster and CAP3 programs and producing putative

transcripts. From the CAP3 output, ESTpass detects chimeric EST sequences which

are confirmed through comparison with the nr database. The last step is

annotating the putative transcript sequences using RefSeq, InterPro, GO and KEGG

gene databases according to user-specified options. The major advantages of

ESTpass are the integration of cleansing and annotating processes, rigorous

chimeric EST detection, exhaustive annotation, and email reporting to inform the

user about the progress and to send the analysis results. The ESTpass results

include three reports (summary, cleansing and annotation) and download function,

as well as graphic statistics. They can be retrieved and downloaded using a

standard web browser. The server is available at http://estpass.kobic.re.kr/.

DOI: 10.1093/nar/gkm369

PMCID: PMC1933161

PMID: 17526512 [Indexed for MEDLINE]

2679. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W300-4. Epub 2007 May 21.

RADAR: a web server for RNA data analysis and research.

Khaladkar M(1), Bellofatto V, Wang JT, Tian B, Shapiro BA.

Author information:

(1)Bioinformatics Program and Department of Computer Science, New Jersey

Institute of Technology, NJ 07102, USA.

RADAR is a web server that provides a multitude of functionality for RNA data

analysis and research. It can align structure-annotated RNA sequences so that

both sequence and structure information are taken into consideration during the

alignment process. This server is capable of performing pairwise structure

alignment, multiple structure alignment, database search and clustering. In

addition, RADAR provides two salient features: (i) constrained alignment of RNA

secondary structures, and (ii) prediction of the consensus structure for a set of

RNA sequences. RADAR will be able to assist scientists in performing many

important RNA mining operations, including the understanding of the functionality

of RNA sequences, the detection of RNA structural motifs and the clustering of

RNA molecules, among others. The web server together with a software package for

download is freely accessible at

http://datalab.njit.edu/biodata/rna/RSmatch/server.htm and

http://www.ccrnp.ncifcrf.gov/~bshapiro/

DOI: 10.1093/nar/gkm253

PMCID: PMC1933136

PMID: 17517784 [Indexed for MEDLINE]

2680. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W653-8. Epub 2007 May 21.

COMPASS server for remote homology inference.

Sadreyev RI(1), Tang M, Kim BH, Grishin NV.

Author information:

(1)Howard Hughes Medical Institute, University of Texas Southwestern Medical

Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA.

sadreyev@chop.swmed.edu

COMPASS is a method for homology detection and local alignment construction based

on the comparison of multiple sequence alignments (MSAs). The method derives

numerical profiles from given MSAs, constructs local profile-profile alignments

and analytically estimates E-values for the detected similarities. Until now,

COMPASS was only available for download and local installation. Here, we present

a new web server featuring the latest version of COMPASS, which provides (i)

increased sensitivity and selectivity of homology detection; (ii) longer, more

complete alignments; and (iii) faster computational speed. After submission of

the query MSA or single sequence, the server performs searches versus a

user-specified database. The server includes detailed and intuitive control of

the search parameters. A flexible output format, structured similarly to BLAST

and PSI-BLAST, provides an easy way to read and analyze the detected profile

similarities. Brief help sections are available for all input parameters and

output options, along with detailed documentation. To illustrate the value of

this tool for protein structure-functional prediction, we present two examples of

detecting distant homologs for uncharacterized protein families. Available at

http://prodata.swmed.edu/compass.

DOI: 10.1093/nar/gkm293

PMCID: PMC1933213

PMID: 17517780 [Indexed for MEDLINE]

2681. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W86-90. Epub 2007 May 21.

MAGMA: analysis of two-channel microarrays made easy.

Rehrauer H(1), Zoller S, Schlapbach R.

Author information:

(1)Functional Genomics Center Zurich, UZH/ETH Zurich, Winterthurerstrasse 190,

8057 Zurich, Switzerland. Hubert.Rehrauer@fgcz.uzh.ch

The web application MAGMA provides a simple and intuitive interface to identify

differentially expressed genes from two-channel microarray data. While the

underlying algorithms are not superior to those of similar web applications,

MAGMA is particularly user friendly and can be used without prior training. The

user interface guides the novice user through the most typical microarray

analysis workflow consisting of data upload, annotation, normalization and

statistical analysis. It automatically generates R-scripts that document MAGMA's

entire data processing steps, thereby allowing the user to regenerate all results

in his local R installation. The implementation of MAGMA follows the

model-view-controller design pattern that strictly separates the R-based

statistical data processing, the web-representation and the application logic.

This modular design makes the application flexible and easily extendible by

experts in one of the fields: statistical microarray analysis, web design or

software development. State-of-the-art Java Server Faces technology was used to

generate the web interface and to perform user input processing. MAGMA's

object-oriented modular framework makes it easily extendible and applicable to

other fields and demonstrates that modern Java technology is also suitable for

rather small and concise academic projects. MAGMA is freely available at

www.magma-fgcz.uzh.ch.

DOI: 10.1093/nar/gkm302

PMCID: PMC1933123

PMID: 17517778 [Indexed for MEDLINE]

2682. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W425-8. Epub 2007 May 21.

Protein knot server: detection of knots in protein structures.

Kolesov G(1), Virnau P, Kardar M, Mirny LA.

Author information:

(1)Harvard-MIT Division of Health Sciences and Technology, Massachusetts

Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA.

KNOTS (http://knots.mit.edu) is a web server that detects knots in protein

structures. Several protein structures have been reported to contain intricate

knots. The physiological role of knots and their effect on folding and evolution

is an area of active research. The user submits a PDB id or uploads a 3D protein

structure in PDB or mmCIF format. The current implementation of the server uses

the Alexander polynomial to detect knots. The results of the analysis that are

presented to the user are the location of the knot in the structure, the type of

the knot and an interactive visualization of the knot. The results can also be

downloaded and viewed offline. The server also maintains a regularly updated list

of known knots in protein structures.

DOI: 10.1093/nar/gkm312

PMCID: PMC1933242

PMID: 17517776 [Indexed for MEDLINE]

2683. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W713-7. Epub 2007 May 21.

DSHIFT: a web server for predicting DNA chemical shifts.

Lam SL(1).

Author information:

(1)Department of Chemistry, The Chinese University of Hong Kong, Shatin, New

Territories, Hong Kong. lams@cuhk.edu.hk

DSHIFT is a web server for predicting chemical shifts of DNA sequences in random

coil form or double helical B-form. The prediction methods are based on sets of

published reference chemical shift values and correction factors which account

for shielding or deshielding effects from neighboring nucleotides. Proton, carbon

and phosphorus chemical shift predictions are available for random coil DNAs. For

double helical B-DNA, only proton chemical shift prediction is available. Results

from these predictions will be useful for facilitating NMR resonance assignments

and investigating structural features of solution DNA molecules. The URL of this

server is: http://www.chem.cuhk.edu.hk/DSHIFT.

DOI: 10.1093/nar/gkm320

PMCID: PMC1933157

PMID: 17517771 [Indexed for MEDLINE]

2684. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W588-94. Epub 2007 May 21.

KinasePhos 2.0: a web server for identifying protein kinase-specific

phosphorylation sites based on sequences and coupling patterns.

Wong YH(1), Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, Chu CH, Huang HD, Ko

MT, Hwang JK.

Author information:

(1)Institute of Bioinformatics, National Chiao Tung University, Hsin-chu 300,

Taiwan.

Due to the importance of protein phosphorylation in cellular control, many

researches are undertaken to predict the kinase-specific phosphorylation sites.

Referred to our previous work, KinasePhos 1.0, incorporated profile hidden Markov

model (HMM) with flanking residues of the kinase-specific phosphorylation sites.

Herein, a new web server, KinasePhos 2.0, incorporates support vector machines

(SVM) with the protein sequence profile and protein coupling pattern, which is a

novel feature used for identifying phosphorylation sites. The coupling pattern

[XdZ] denotes the amino acid coupling-pattern of amino acid types X and Z that

are separated by d amino acids. The differences or quotients of coupling strength

C(XdZ) between the positive set of phosphorylation sites and the background set

of whole protein sequences from Swiss-Prot are computed to determine the number

of coupling patterns for training SVM models. After the evaluation based on

k-fold cross-validation and Jackknife cross-validation, the average predictive

accuracy of phosphorylated serine, threonine, tyrosine and histidine are 90, 93,

88 and 93%, respectively. KinasePhos 2.0 performs better than other tools

previously developed. The proposed web server is freely available at

http://KinasePhos2.mbc.nctu.edu.tw/.

DOI: 10.1093/nar/gkm322

PMCID: PMC1933228

PMID: 17517770 [Indexed for MEDLINE]

2685. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W115-20. Epub 2007 May 21.

QuickSNP: an automated web server for selection of tagSNPs.

Grover D(1), Woodfield AS, Verma R, Zandi PP, Levinson DF, Potash JB.

Author information:

(1)Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of

Medicine, Baltimore, MD 21287, USA.

Although large-scale genetic association studies involving hundreds to thousands

of SNPs have become feasible, the associated cost is substantial. Even with the

increased efficiency introduced by the use of tagSNPs, researchers are often

seeking ways to maximize resource utilization given a set of SNP-based

gene-mapping goals. We have developed a web server named QuickSNP in order to

provide cost-effective selection of SNPs, and to fill in some of the gaps in

existing SNP selection tools. One useful feature of QuickSNP is the option to

select only gene-centric SNPs from a chromosomal region in an automated fashion.

Other useful features include automated selection of coding non-synonymous SNPs,

SNP filtering based on inter-SNP distances and information regarding the

availability of genotyping assays for SNPs and whether they are present on whole

genome chips. The program produces user-friendly summary tables and results, and

a link to a UCSC Genome Browser track illustrating the position of the selected

tagSNPs in relation to genes and other genomic features. We hope the unique

combination of features of this server will be useful for researchers aiming to

select markers for their genotyping studies. The server is freely available and

can be accessed at the URL http://bioinformoodics.jhmi.edu/quickSNP.pl.

DOI: 10.1093/nar/gkm329

PMCID: PMC1933212

PMID: 17517769 [Indexed for MEDLINE]

2686. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W363-8. Epub 2007 May 21.

M4T: a comparative protein structure modeling server.

Fernandez-Fuentes N(1), Madrid-Aliste CJ, Rai BK, Fajardo JE, Fiser A.

Author information:

(1)Department of Biochemistry and Seaver Foundation Center for Bioinformatics,

Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461,

USA.

Multiple Mapping Method with Multiple Templates (M4T)

(http://www.fiserlab.org/servers/m4t) is a fully automated comparative protein

structure modeling server. The novelty of M4T resides in two of its major

modules, Multiple Templates (MT) and Multiple Mapping Method (MMM). The MT module

of M4T selects and optimally combines the sequences of multiple template

structures through an iterative clustering approach that takes into account the

'unique' contribution of each template, its sequence similarity to other template

sequences and to the target sequences, and the quality of its experimental

resolution. MMM module is a sequence-to-structure alignment method that is aimed

at improving the alignment accuracy, especially at lower sequence identity

levels. The current implementation of MMM takes inputs from three

profile-to-profile-based alignment methods and iteratively compares and ranks

alternatively aligned regions according to their fit in the structural

environment of the template structure. The performance of M4T was benchmarked on

CASP6 comparative modeling target sequences and on a larger independent test set

and showed a favorable performance to current state-of-the-art methods.

DOI: 10.1093/nar/gkm341

PMCID: PMC1933164

PMID: 17517764 [Indexed for MEDLINE]

2687. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W561-7. Epub 2007 May 21.

3D-partner: a web server to infer interacting partners and binding models.

Chen YC(1), Lo YS, Hsu WC, Yang JM.

Author information:

(1)Institute of Bioinformatics, National Chiao Tung University, Hsinchu, 30050,

Taiwan.

The 3D-partner is a web tool to predict interacting partners and binding models

of a query protein sequence through structure complexes and a new scoring

function. 3D-partner first utilizes IMPALA to identify homologous structures

(templates) of a query from a heterodimer profile library. The

interacting-partner sequence profiles of these templates are then used to search

interacting candidates of the query from protein sequence databases (e.g.

SwissProt) by PSI-BLAST. We developed a new scoring function, which includes the

contact-residue interacting score (e.g. the steric, hydrogen bonds, and

electrostatic interactions) and the template consensus score (e.g.

couple-conserved residue and the template similarity scores), to evaluate how

well the interfaces between the query and interacting candidates. Based on this

scoring function, 3D-partner provides the statistic significance, the binding

models (e.g. hydrogen bonds and conserved amino acids) and functional annotations

of interacting partners. The correlation between experimental energies and

predicted binding affinities of our scoring function is 0.91 on 275 mutated

residues from the ASEdb. The average precision of the server is 0.72 on 563

queries and the execution time of this server for a query is approximately 15 s

on average. These results suggest that the 3D-partner server can be useful in

protein-protein interaction predictions and binding model visualizations. The

server is available online at: http://3D-partner.life.nctu.edu.tw.

DOI: 10.1093/nar/gkm346

PMCID: PMC1933210

PMID: 17517763 [Indexed for MEDLINE]

2688. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W633-8. Epub 2007 May 21.

INDELSCAN: a web server for comparative identification of species-specific and

non-species-specific insertion/deletion events.

Chen FC(1), Chen CJ, Chuang TJ.

Author information:

(1)Division of Biostatistics and Bioinformatics, National Health Research

Institute, Miaoli County 350, Taiwan.

Insertion and deletion (indel) events usually have dramatic effects on genome

structure and gene function. Species-specific indels have been demonstrated to be

associated with species-unique traits. Currently, indel identifications mainly

rely on pair-wise sequence alignments (the 'pair-wise indels'), which suffer lack

of discrimination of species specificity and insertion versus deletion. Also,

there is no freely accessible web server for genome-wide identification of

indels. Therefore, we develop a web server--INDELSCAN--to identify four types of

indels using multiple sequence alignments that include sequences from one target,

one subject and > or =1 out-group species. The four types of indels identified

encompass target species-specific, subject species-specific, non-species-specific

and target-subject pair-wise indels. Insertions and deletions are discriminated

with reference to out-group sequences. The genomic locations (5'UTR, intron, CDS,

3'UTR and intergenic region) of these indels are also provided for functional

analysis. INDELSCAN provides genomic sequences and gene annotations from a wide

spectrum of taxa for users to select from, including nine target species (human

(Homo sapiens), mouse (Mus musculus), rat (Rattus norvegicus), dog (Canis

familiaris), opossum (Monodelphis domestica), chicken (Gallus gallus), zebrafish

(Danio rerio), fly (Drosophila melanogaster) and yeast (Saccharomyces cerevisiae)

and >35 subject/out-group species, ranging from yeasts to mammals. The server

also provides analytic figures and supports indel identification from

user-uploaded alignments/annotations. INDELSCAN is freely accessible at

http://indelscan.genomics.sinica.edu.tw/IndelScan/.

DOI: 10.1093/nar/gkm350

PMCID: PMC1933116

PMID: 17517762 [Indexed for MEDLINE]

2689. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W416-9. Epub 2007 May 8.

3dLOGO: a web server for the identification, analysis and use of conserved

protein substructures.

Via A(1), Peluso D, Gherardini PF, de Rinaldis E, Colombo T, Ausiello G,

Helmer-Citterich M.

Author information:

(1)Centre for Molecular Bioinformatics, Department of Biology, University of Rome

Tor Vergata, 00133 Rome, Italy. allegra.via@uniroma2.it

3dLOGO is a web server for the identification and analysis of conserved protein

3D substructures. Given a set of residues in a PDB (Protein Data Bank) chain, the

server detects the matching substructure(s) in a set of user-provided protein

structures, generates a multiple structure alignment centered on the input

substructures and highlights other residues whose structural conservation becomes

evident after the defined superposition. Conserved residues are proposed to the

user for highlighting functional areas, deriving refined structural motifs or

building sequence patterns. Residue structural conservation can be visualized

through an expressly designed Java application, 3dProLogo, which is a 3D

implementation of a sequence logo. The 3dLOGO server, with related documentation,

is available at http://3dlogo.uniroma2.it/

DOI: 10.1093/nar/gkm228

PMCID: PMC1933223

PMID: 17488847 [Indexed for MEDLINE]

2690. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W522-5. Epub 2007 May 8.

PDB2PQR: expanding and upgrading automated preparation of biomolecular structures

for molecular simulations.

Dolinsky TJ(1), Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA.

Author information:

(1)Department of Biochemistry and Molecular Biophysics, Center for Computational

Biology, Washington University in St. Louis, 700 S. Euclid Ave., Campus Box 8036,

St. Louis, MO 63110, USA.

Real-world observable physical and chemical characteristics are increasingly

being calculated from the 3D structures of biomolecules. Methods for calculating

pK(a) values, binding constants of ligands, and changes in protein stability are

readily available, but often the limiting step in computational biology is the

conversion of PDB structures into formats ready for use with biomolecular

simulation software. The continued sophistication and integration of biomolecular

simulation methods for systems- and genome-wide studies requires a fast, robust,

physically realistic and standardized protocol for preparing macromolecular

structures for biophysical algorithms. As described previously, the PDB2PQR web

server addresses this need for electrostatic field calculations (Dolinsky et al.,

Nucleic Acids Research, 32, W665-W667, 2004). Here we report the significantly

expanded PDB2PQR that includes the following features: robust standalone command

line support, improved pK(a) estimation via the PROPKA framework, ligand

parameterization via PEOE\_PB charge methodology, expanded set of force fields and

easily incorporated user-defined parameters via XML input files, and improvement

of atom addition and optimization code. These features are available through a

new web interface (http://pdb2pqr.sourceforge.net/), which offers users a wide

range of options for PDB file conversion, modification and parameterization.

DOI: 10.1093/nar/gkm276

PMCID: PMC1933214

PMID: 17488841 [Indexed for MEDLINE]

2691. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W543-8. Epub 2007 May 8.

ProMateus--an open research approach to protein-binding sites analysis.

Neuvirth H(1), Heinemann U, Birnbaum D, Tishby N, Schreiber G.

Author information:

(1)School of Computer Science and Engineering, The Hebrew University Jerusalem,

91904, Israel.

The development of bioinformatic tools by individual labs results in the

abundance of parallel programs for the same task. For example, identification of

binding site regions between interacting proteins is done using: ProMate, WHISCY,

PPI-Pred, PINUP and others. All servers first identify unique properties of

binding sites and then incorporate them into a predictor. Obviously, the

resulting prediction would improve if the most suitable parameters from each of

those predictors would be incorporated into one server. However, because of the

variation in methods and databases, this is currently not feasible. Here, the

protein-binding site prediction server is extended into a general protein-binding

sites research tool, ProMateus. This web tool, based on ProMate's infrastructure

enables the easy exploration and incorporation of new features and databases by

the user, providing an evaluation of the benefit of individual features and their

combination within a set framework. This transforms the individual research into

a community exercise, bringing out the best from all users for optimized

predictions. The analysis is demonstrated on a database of protein protein and

protein-DNA interactions. This approach is basically different from that used in

generating meta-servers. The implications of the open-research approach are

discussed. ProMateus is available at http://bip.weizmann.ac.il/promate.

DOI: 10.1093/nar/gkm301

PMCID: PMC1933218

PMID: 17488838 [Indexed for MEDLINE]

2692. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W325-9. Epub 2007 May 8.

taveRNA: a web suite for RNA algorithms and applications.

Aksay C(1), Salari R, Karakoc E, Alkan C, Sahinalp SC.

Author information:

(1)Lab for Computational Biology, SFU, Canada.

We present taveRNA, a web server package that hosts three RNA web services:

alteRNA, inteRNA and pRuNA. alteRNA is a new alternative for RNA secondary

structure prediction. It is based on a dynamic programming solution that

minimizes the sum of energy density and free energy of an RNA structure. inteRNA

is the first RNA-RNA interaction structure prediction web service. It also

employs a dynamic programming algorithm to minimize the free energy of the

resulting joint structure of the two interacting RNAs. Lastly, pRuNA is an

efficient database pruning service; which given a query RNA, eliminates a

significant portion of an ncRNA database and returns only a few ncRNAs as

potential regulators. taveRNA is available at http://compbio.cs.sfu.ca/taverna.

DOI: 10.1093/nar/gkm303

PMCID: PMC1933159

PMID: 17488837 [Indexed for MEDLINE]

2693. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W66-70. Epub 2007 May 8.

NTMG (N-terminal Truncated Mutants Generator for cDNA): an automatic multiplex

PCR assays design for generating various N-terminal truncated cDNA mutants.

Chen YF(1), Chen RC, Tseng LY, Lin E, Chan YK, Pan RH.

Author information:

(1)Department of Health Services Management, China Medical University, Taichung,

Taiwan, ROC.

The sequential deletion method is generally used to locate the functional domain

of a protein. With this method, in order to find the various N-terminal truncated

mutants, researchers have to investigate the ATG-like codons, to design various

multiplex polymerase chain reaction (PCR) forward primers and to do several PCR

experiments. This web server (N-terminal Truncated Mutants Generator for cDNA)

will automatically generate groups of forward PCR primers and the corresponding

reverse PCR primers that can be used in a single batch of a multiplex PCR

experiment to extract the various N-terminal truncated mutants. This saves much

time and money for those who use the sequential deletion method in their

research. This server is available at http://oblab.cs.nchu.edu.tw:8080/WebSDL/.

DOI: 10.1093/nar/gkm305

PMCID: PMC1933230

PMID: 17488836 [Indexed for MEDLINE]

2694. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W27-32. Epub 2007 May 8.

Berkeley Phylogenomics Group web servers: resources for structural phylogenomic

analysis.

Glanville JG(1), Kirshner D, Krishnamurthy N, Sjölander K.

Author information:

(1)Berkeley Phylogenomics Group, University of California, Berkeley, USA.

Phylogenomic analysis addresses the limitations of function prediction based on

annotation transfer, and has been shown to enable the highest accuracy in

prediction of protein molecular function. The Berkeley Phylogenomics Group

provides a series of web servers for phylogenomic analysis: classification of

sequences to pre-computed families and subfamilies using the PhyloFacts

Phylogenomic Encyclopedia, FlowerPower clustering of proteins sharing the same

domain architecture, MUSCLE multiple sequence alignment, SATCHMO simultaneous

alignment and tree construction and SCI-PHY subfamily identification. The

PhyloBuilder web server provides an integrated phylogenomic pipeline starting

with a user-supplied protein sequence, proceeding to homolog identification,

multiple alignment, phylogenetic tree construction, subfamily identification and

structure prediction. The Berkeley Phylogenomics Group resources are available at

http://phylogenomics.berkeley.edu.

DOI: 10.1093/nar/gkm325

PMCID: PMC1933202

PMID: 17488835 [Indexed for MEDLINE]

2695. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W675-7. Epub 2007 May 7.

eProbalign: generation and manipulation of multiple sequence alignments using

partition function posterior probabilities.

Chikkagoudar S(1), Roshan U, Livesay D.

Author information:

(1)Department of Computer Science, New Jersey Institute of Technology, University

of North Carolina at Charlotte, USA.

Probalign computes maximal expected accuracy multiple sequence alignments from

partition function posterior probabilities. To date, Probalign is among the very

best scoring methods on the BAliBASE, HOMSTRAD and OXBENCH benchmarks. Here, we

introduce eProbalign, which is an online implementation of the approach.

Moreover, the eProbalign web server doubles as an online platform for

post-alignment analysis. The heart-and-soul of the post-alignment functionality

is the Probalign Alignment Viewer applet, which provides users a convenient means

to manipulate the alignments by posterior probabilities. The viewer can also be

used to produce graphical and text versions of the output. The eProbalign web

server and underlying Probalign source code is freely accessible at

http://probalign.njit.edu.

DOI: 10.1093/nar/gkm267

PMCID: PMC1933135

PMID: 17485479 [Indexed for MEDLINE]

2696. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W265-8. Epub 2007 May 7.

LTR\_FINDER: an efficient tool for the prediction of full-length LTR

retrotransposons.

Xu Z(1), Wang H.

Author information:

(1)T-Life Research Center, Fudan University, 220 HanDan Road, Shanghai, 200433,

China.

Long terminal repeat retrotransposons (LTR elements) are ubiquitous eukaryotic

transposable elements. They play important roles in the evolution of genes and

genomes. Ever-growing amount of genomic sequences of many organisms present a

great challenge to fast identifying them. That is the first and indispensable

step to study their structure, distribution, functions and other biological

impacts. However, until today, tools for efficient LTR retrotransposon discovery

are very limited. Thus, we developed LTR\_FINDER web server. Given DNA sequences,

it predicts locations and structure of full-length LTR retrotransposons

accurately by considering common structural features. LTR\_FINDER is a system

capable of scanning large-scale sequences rapidly and the first web server for ab

initio LTR retrotransposon finding. We illustrate its usage and performance on

the genome of Saccharomyces cerevisiae. The web server is freely accessible at

http://tlife.fudan.edu.cn/ltr\_finder/.

DOI: 10.1093/nar/gkm286

PMCID: PMC1933203

PMID: 17485477 [Indexed for MEDLINE]

2697. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W438-43. Epub 2007 May 7.

fastSCOP: a fast web server for recognizing protein structural domains and SCOP

superfamilies.

Tung CH(1), Yang JM.

Author information:

(1)Institute of Bioinformatics, National Chiao Tung University, Hsinchu, 30050

Taiwan.

The fastSCOP is a web server that rapidly identifies the structural domains and

determines the evolutionary superfamilies of a query protein structure. This

server uses 3D-BLAST to scan quickly a large structural classification database

(SCOP1.71 with <95% identity with each other) and the top 10 hit domains, which

have different superfamily classifications, are obtained from the hit lists.

MAMMOTH, a detailed structural alignment tool, is adopted to align these top 10

structures to refine domain boundaries and to identify evolutionary

superfamilies. Our previous works demonstrated that 3D-BLAST is as fast as BLAST,

and has the characteristics of BLAST (e.g. a robust statistical basis, effective

search and reliable database search capabilities) in large structural database

searches based on a structural alphabet database and a structural alphabet

substitution matrix. The classification accuracy of this server is approximately

98% for 586 query structures and the average execution time is approximately 5.

This server was also evaluated on 8700 structures, which have no annotations in

the SCOP; the server can automatically assign 7311 (84%) proteins (9420 domains)

to the SCOP superfamilies in 9.6 h. These results suggest that the fastSCOP is

robust and can be a useful server for recognizing the evolutionary

classifications and the protein functions of novel structures. The server is

accessible at http://fastSCOP.life.nctu.edu.tw.

DOI: 10.1093/nar/gkm288

PMCID: PMC1933144

PMID: 17485476 [Indexed for MEDLINE]

2698. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W451-4. Epub 2007 May 7.

SH3-Hunter: discovery of SH3 domain interaction sites in proteins.

Ferraro E(1), Peluso D, Via A, Ausiello G, Helmer-Citterich M.

Author information:

(1)Centre for Molecular Bioinformatics, Department of Biology, University of Rome

Tor Vergata, Via della Ricerca Scientifica, 00133, Rome, Italy.

enrico.ferraro@uniroma2.it

SH3-Hunter (http://cbm.bio.uniroma2.it/SH3-Hunter/) is a web server for the

recognition of putative SH3 domain interaction sites on protein sequences. Given

an input query consisting of one or more protein sequences, the server identifies

peptides containing poly-proline binding motifs and associates them to a list of

SH3 domains, in order to compose peptide-domain pairs. The server can accept a

list of peptides and allows users to upload an input file in a proper format. An

accurate selection of SH3 domains is available and users can also submit their

own SH3 domain sequence. SH3-Hunter evaluates which peptide-domain pair

represents a possible interaction pair and produces as output a list of

significant interaction sites for each query protein. Each proposed interaction

site is associated to a propensity score and sensitivity and precision levels for

the prediction. The server prediction capability is based on a neural network

model integrating high-throughput pep-spot data with structural information

extracted from known SH3-peptide complexes.

DOI: 10.1093/nar/gkm296

PMCID: PMC1933191

PMID: 17485474 [Indexed for MEDLINE]

2699. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W259-64. Epub 2007 May 7.

SCOPE: a web server for practical de novo motif discovery.

Carlson JM(1), Chakravarty A, DeZiel CE, Gross RH.

Author information:

(1)Department of Computer Science and Engineering, University of Washington,

Seattle, WA, USA.

SCOPE is a novel parameter-free method for the de novo identification of

potential regulatory motifs in sets of coordinately regulated genes. The SCOPE

algorithm combines the output of three component algorithms, each designed to

identify a particular class of motifs. Using an ensemble learning approach, SCOPE

identifies the best candidate motifs from its component algorithms. In tests on

experimentally determined datasets, SCOPE identified motifs with a significantly

higher level of accuracy than a number of other web-based motif finders run with

their default parameters. Because SCOPE has no adjustable parameters, the web

server has an intuitive interface, requiring only a set of gene names or FASTA

sequences and a choice of species. The most significant motifs found by SCOPE are

displayed graphically on the main results page with a table containing summary

statistics for each motif. Detailed motif information, including the sequence

logo, PWM, consensus sequence and specific matching sites can be viewed through a

single click on a motif. SCOPE's efficient, parameter-free search strategy has

enabled the development of a web server that is readily accessible to the

practising biologist while providing results that compare favorably with those of

other motif finders. The SCOPE web server is at

<http://genie.dartmouth.edu/scope>.

DOI: 10.1093/nar/gkm310

PMCID: PMC1933170

PMID: 17485471 [Indexed for MEDLINE]

2700. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W531-7. Epub 2007 May 7.

The RCI server: rapid and accurate calculation of protein flexibility using

chemical shifts.

Berjanskii MV(1), Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB, Canada

T6G 2E8.

Protein motions play important roles in numerous biological processes such as

enzyme catalysis, muscle contractions, antigen-antibody interactions, gene

regulation and virus assembly. Knowledge of protein flexibility is also important

in rational drug design, protein docking and protein engineering. However, the

experimental measurement of protein motions is often difficult, requiring

sophisticated experiments, complex data analysis and detailed information about

the protein's tertiary structure. As a result, there is a considerable interest

in developing simpler, more effective ways of quantifying protein flexibility.

Recently, we described a method, called the random coil index (RCI), which is

able to quantitatively estimate backbone root mean square fluctuations (RMSFs) of

structural ensembles and order parameters using only chemical shifts. The RCI

method is very fast (<5 s) and exceedingly robust. It also offers an excellent

alternative to traditional methods of measuring protein flexibility. We have

recently extended the RCI concept and implemented it as a web server. This server

allows facile, accurate and fully automated predictions of MD RMSF values, NMR

RMSF values and model-free order parameters (S2) directly from chemical shift

assignments. It also performs automatic chemical shift re-referencing to ensure

consistency and reproducibility. On average, the correlation between RCI

predictions and experimentally obtained motional amplitudes is within the range

from 0.77 to 0.82. The server is available at

http://wishart.biology.ualberta.ca/rci.

DOI: 10.1093/nar/gkm328

PMCID: PMC1933179

PMID: 17485469 [Indexed for MEDLINE]

2701. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W429-32. Epub 2007 May 5.

Advantages of combined transmembrane topology and signal peptide prediction--the

Phobius web server.

Käll L(1), Krogh A, Sonnhammer EL.

Author information:

(1)Center for Genomics and Bioinformatics, Karolinska Institutet, S-17177

Stockholm, Sweden. lukall@u.washington.edu

When using conventional transmembrane topology and signal peptide predictors,

such as TMHMM and SignalP, there is a substantial overlap between these two types

of predictions. Applying these methods to five complete proteomes, we found that

30-65% of all predicted signal peptides and 25-35% of all predicted transmembrane

topologies overlap. This impairs predictions of 5-10% of the proteome, hence this

is an important issue in protein annotation. To address this problem, we

previously designed a hidden Markov model, Phobius, that combines transmembrane

topology and signal peptide predictions. The method makes an optimal choice

between transmembrane segments and signal peptides, and also allows constrained

and homology-enriched predictions. We here present a web interface

(http://phobius.cgb.ki.se and http://phobius.binf.ku.dk) to access Phobius.

DOI: 10.1093/nar/gkm256

PMCID: PMC1933244

PMID: 17483518 [Indexed for MEDLINE]

2702. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W613-8. Epub 2007 May 5.

MetaPath Online: a web server implementation of the network expansion algorithm.

Handorf T(1), Ebenhöh O.

Author information:

(1)Department of Theoretical Biophysics, Institute of Biology, Humboldt

University Berlin, Invalidenstrasse 42, 10115 Berlin, Germany.

We designed a web server for the analysis of biosynthetic capacities of metabolic

networks. The implementation is based on the network expansion algorithm and the

concept of scopes. For a given network and predefined external resources, called

the seed metabolites, the scope is defined as the set of products which the

network is in principle able to produce. Through the web interface the user can

select a variety of metabolic networks or provide his or her own list of

reactions. The information on the organism-specific networks has been extracted

from the KEGG database. By choosing an arbitrary set of seed compounds, the user

can obtain the corresponding scopes. With our web server application we provide

an easy to use interface to perform a variety of structural and functional

network analyses. Problems that can be addressed using the web server include the

calculation of synthesizing capacities, the visualization of synthesis pathways,

functional analysis of mutant networks or comparative analysis of related

species. The web server is accessible through

http://scopes.biologie.hu-berlin.de.

DOI: 10.1093/nar/gkm287

PMCID: PMC1933239

PMID: 17483511 [Indexed for MEDLINE]

2703. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W578-84. Epub 2007 May 5.

RNABindR: a server for analyzing and predicting RNA-binding sites in proteins.

Terribilini M(1), Sander JD, Lee JH, Zaback P, Jernigan RL, Honavar V, Dobbs D.

Author information:

(1)Department of Genetics, Development & Cell Biology, Bioinformatics &

Computational Biology Program, Iowa State University, Ames, Iowa 50011, USA.

Understanding interactions between proteins and RNA is key to deciphering the

mechanisms of many important biological processes. Here we describe RNABindR, a

web-based server that identifies and displays RNA-binding residues in known

protein-RNA complexes and predicts RNA-binding residues in proteins of unknown

structure. RNABindR uses a distance cutoff to identify which amino acids contact

RNA in solved complex structures (from the Protein Data Bank) and provides a

labeled amino acid sequence and a Jmol graphical viewer in which RNA-binding

residues are displayed in the context of the three-dimensional structure.

Alternatively, RNABindR can use a Naive Bayes classifier trained on a

non-redundant set of protein-RNA complexes from the PDB to predict which amino

acids in a protein sequence of unknown structure are most likely to bind RNA.

RNABindR automatically displays 'high specificity' and 'high sensitivity'

predictions of RNA-binding residues. RNABindR is freely available at

http://bindr.gdcb.iastate.edu/RNABindR.

DOI: 10.1093/nar/gkm294

PMCID: PMC1933119

PMID: 17483510 [Indexed for MEDLINE]

2704. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W47-51. Epub 2007 May 3.

RF-DYMHC: detecting the yeast meiotic recombination hotspots and coldspots by

random forest model using gapped dinucleotide composition features.

Jiang P(1), Wu H, Wei J, Sang F, Sun X, Lu Z.

Author information:

(1)State Key Laboratory of Bioelectronics, Department of Biological Science and

Medical Engineering, Southeast University, Nanjing, 210096, PR China.

In the yeast, meiotic recombination is initiated by double-strand DNA breaks

(DSBs) which occur at relatively high frequencies in some genomic regions

(hotspots) and relatively low frequencies in others (coldspots). Although

observations concerning individual hot/cold spots have given clues as to the

mechanism of recombination initiation, the prediction of hot/cold spots from DNA

sequence information is a challenging task. In this article, we introduce a

random forest (RF) prediction model to detect recombination hot/cold spots from

yeast genome. The out-of-bag (OOB) estimation of the model indicated that the RF

classifier achieved high prediction performance with 82.05% total accuracy and

0.638 Mattew's correlation coefficient (MCC) value. Compared with an alternative

machine-learning algorithm, support vector machine (SVM), the RF method

outperforms it in both sensitivity and specificity. The prediction model is

implemented as a web server (RF-DYMHC) and it is freely available at

http://www.bioinf.seu.edu.cn/Recombination/rf\_dymhc.htm. Given a yeast genome and

prediction parameters (RI-value and non-overlapping window scan size), the

program reports the predicted hot/cold spots and marks them in color.

DOI: 10.1093/nar/gkm217

PMCID: PMC1933199

PMID: 17478517 [Indexed for MEDLINE]

2705. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W212-6. Epub 2007 May 3.

Update of the G2D tool for prioritization of gene candidates to inherited

diseases.

Perez-Iratxeta C(1), Bork P, Andrade-Navarro MA.

Author information:

(1)Ontario Genomics Innovation Centre, Ottawa Health Research Institute, 501

Smyth, Ottawa, ON, Canada K1H 8L6. cperez-iratxeta@ohri.ca

G2D (genes to diseases) is a web resource for prioritizing genes as candidates

for inherited diseases. It uses three algorithms based on different

prioritization strategies. The input to the server is the genomic region where

the user is looking for the disease-causing mutation, plus an additional piece of

information depending on the algorithm used. This information can either be the

disease phenotype (described as an online Mendelian inheritance in man (OMIM)

identifier), one or several genes known or suspected to be associated with the

disease (defined by their Entrez Gene identifiers), or a second genomic region

that has been linked as well to the disease. In the latter case, the tool uses

known or predicted interactions between genes in the two regions extracted from

the STRING database. The output in every case is an ordered list of candidate

genes in the region of interest. For the first two of the three methods, the

candidate genes are first retrieved through sequence homology search, then scored

accordingly to the corresponding method. This means that some of them will

correspond to well-known characterized genes, and others will overlap with

predicted genes, thus providing a wider analysis. G2D is publicly available at

http://www.ogic.ca/projects/g2d\_2/

DOI: 10.1093/nar/gkm223

PMCID: PMC1933178

PMID: 17478516 [Indexed for MEDLINE]

2706. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W193-200. Epub 2007 May 3.

g:Profiler--a web-based toolset for functional profiling of gene lists from

large-scale experiments.

Reimand J(1), Kull M, Peterson H, Hansen J, Vilo J.

Author information:

(1)Institute of Computer Science, University of Tartu, Liivi 2, 50409 Tartu,

Estonia.

g:Profiler (http://biit.cs.ut.ee/gprofiler/) is a public web server for

characterising and manipulating gene lists resulting from mining high-throughput

genomic data. g:Profiler has a simple, user-friendly web interface with powerful

visualisation for capturing Gene Ontology (GO), pathway, or transcription factor

binding site enrichments down to individual gene levels. Besides standard

multiple testing corrections, a new improved method for estimating the true

effect of multiple testing over complex structures like GO has been introduced.

Interpreting ranked gene lists is supported from the same interface with very

efficient algorithms. Such ordered lists may arise when studying the most

significantly affected genes from high-throughput data or genes co-expressed with

the query gene. Other important aspects of practical data analysis are supported

by modules tightly integrated with g:Profiler. These are: g:Convert for

converting between different database identifiers; g:Orth for finding orthologous

genes from other species; and g:Sorter for searching a large body of public gene

expression data for co-expression. g:Profiler supports 31 different species, and

underlying data is updated regularly from sources like the Ensembl database.

Bioinformatics communities wishing to integrate with g:Profiler can use

alternative simple textual outputs.

DOI: 10.1093/nar/gkm226

PMCID: PMC1933153

PMID: 17478515 [Indexed for MEDLINE]

2707. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W503-5. Epub 2007 May 3.

PAR-3D: a server to predict protein active site residues.

Goyal K(1), Mohanty D, Mande SC.

Author information:

(1)Centre for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad

500 076, India.

PAR-3D (http://sunserver.cdfd.org.in:8080/protease/PAR\_3D/index.html) is a

web-based tool that exploits the fact that relative juxtaposition of active site

residues is a conserved feature in functionally related protein families. The

server uses previously calculated and stored values of geometrical parameters of

a set of known proteins (training set) for prediction of active site residues in

a query protein structure. PAR-3D stores motifs for different classes of

proteases, the ten glycolytic pathway enzymes and metal-binding sites. The server

accepts the structures in the pdb format. The first step during the prediction is

the extraction of probable active site residues from the query structure. Spatial

arrangement of the probable active site residues is then determined in terms of

geometrical parameters. These are compared with stored geometries of the

different motifs. Its speed and efficiency make it a beneficial tool for

structural genomics projects, especially when the biochemical function of the

protein has not been characterized.

DOI: 10.1093/nar/gkm252

PMCID: PMC1933233

PMID: 17478506 [Indexed for MEDLINE]

2708. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W320-4. Epub 2007 May 3.

pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding

windows.

Reeder J(1), Steffen P, Giegerich R.

Author information:

(1)Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany.

RNA pseudoknots are an important structural feature of RNAs, but often neglected

in computer predictions for reasons of efficiency. Here, we present the pknotsRG

Web Server for single sequence RNA secondary structure prediction including

pseudoknots. pknotsRG employs the newest Turner energy rules for finding the

structure of minimal free energy. The algorithm has been improved in several ways

recently. First, it has been reimplemented in the C programming language,

resulting in a 60-fold increase in speed. Second, all suboptimal foldings up to a

user-defined threshold can be enumerated. For large scale analysis, a fast

sliding window mode is available. Further improvements of the Web Server are a

new output visualization using the PseudoViewer Web Service or RNAmovies for a

movie like animation of several suboptimal foldings. The tool is available as

source code, binary executable, online tool or as Web Service. The latter

alternative allows for an easy integration into bio-informatics pipelines.

pknotsRG is available at the Bielefeld Bioinformatics Server

(http://bibiserv.techfak.uni-bielefeld.de/pknotsrg).

DOI: 10.1093/nar/gkm258

PMCID: PMC1933184

PMID: 17478505 [Indexed for MEDLINE]

2709. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W411-5. Epub 2007 May 3.

QSCOP-BLAST--fast retrieval of quantified structural information for protein

sequences of unknown structure.

Suhrer SJ(1), Gruber M, Sippl MJ.

Author information:

(1)Center of Applied Molecular Engineering, Department of Bioinformatics,

University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria.

QSCOP is a quantitative structural classification of proteins which distinguishes

itself from other classifications by two essential properties: (i) QSCOP is

concurrent with the Research Collaboratory for Structural Bioinformatics (RCSB)

Protein Data Bank and (ii) QSCOP covers the widely used SCOP classification with

layers of quantitative structural information. The QSCOP-BLAST web server

presented here combines the BLAST sequence search engine with QSCOP to retrieve,

for a given query sequence, all structural information currently available. The

resulting search engine is reliable in terms of the quality of results obtained,

and it is efficient in that results are displayed instantaneously. The

hierarchical organization of QSCOP is used to control the redundancy and

diversity of the retrieved hits with the benefit that the often cumbersome and

difficult interpretation of search results is an intuitive and straightforward

exercise. We demonstrate the use of QSCOP-BLAST by example. The server is

accessible at http://qscop-blast.services.came.sbg.ac.at/

DOI: 10.1093/nar/gkm264

PMCID: PMC1933160

PMID: 17478501 [Indexed for MEDLINE]

2710. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W549-55. Epub 2007 May 3.

MODPROPEP: a program for knowledge-based modeling of protein-peptide complexes.

Kumar N(1), Mohanty D.

Author information:

(1)National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi 110067,

India.

MODPROPEP is a web server for knowledge-based modeling of protein-peptide

complexes, specifically peptides in complex with major histocompatibility complex

(MHC) proteins and kinases. The available crystal structures of protein-peptide

complexes in PDB are used as templates for modeling peptides of desired sequence

in the substrate-binding pocket of MHCs or protein kinases. The substrate

peptides are modeled using the same backbone conformation as in the template and

the side-chain conformations are obtained by the program SCWRL. MODPROPEP

provides a number of user-friendly interfaces for visualizing the structure of

the modeled protein-peptide complexes and analyzing the contacts made by the

modeled peptide ligand in the substrate-binding pocket of the MHC or protein

kinase. Analysis of these specific inter-molecular contacts is crucial for

understanding structural basis of the substrate specificity of these two protein

families. This software also provides appropriate interfaces for identifying,

putative MHC-binding peptides in the sequence of an antigen or phosphorylation

sites on the substrate protein of a kinase, by scoring these inter-molecular

contacts using residue-based statistical pair potentials. MODPROPEP would

complement various available sequence-based programs (SYFPEITHI, SCANSITE, etc.)

for predicting substrates of MHCs and protein kinases. The program is available

at http://www.nii.res.in/modpropep.html.

DOI: 10.1093/nar/gkm266

PMCID: PMC1933231

PMID: 17478500 [Indexed for MEDLINE]

2711. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W253-8. Epub 2007 May 3.

STAMP: a web tool for exploring DNA-binding motif similarities.

Mahony S(1), Benos PV.

Author information:

(1)Department of Computational Biology, School of Medicine, University of

Pittsburgh, Pittsburgh, PA, USA.

STAMP is a newly developed web server that is designed to support the study of

DNA-binding motifs. STAMP may be used to query motifs against databases of known

motifs; the software aligns input motifs against the chosen database (or

alternatively against a user-provided dataset), and lists of the highest-scoring

matches are returned. Such similarity-search functionality is expected to

facilitate the identification of transcription factors that potentially interact

with newly discovered motifs. STAMP also automatically builds multiple

alignments, familial binding profiles and similarity trees when more than one

motif is inputted. These functions are expected to enable evolutionary studies on

sets of related motifs and fixed-order regulatory modules, as well as

illustrating similarities and redundancies within the input motif collection.

STAMP is a highly flexible alignment platform, allowing users to 'mix-and-match'

between various implemented comparison metrics, alignment methods (local or

global, gapped or ungapped), multiple alignment strategies and tree-building

methods. Motifs may be inputted as frequency matrices (in many of the commonly

used formats), consensus sequences, or alignments of known binding sites. STAMP

also directly accepts the output files from 12 supported motif-finders, enabling

quick interpretation of motif-discovery analyses. STAMP is available at

http://www.benoslab.pitt.edu/stamp.

DOI: 10.1093/nar/gkm272

PMCID: PMC1933206

PMID: 17478497 [Indexed for MEDLINE]

2712. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W81-5. Epub 2007 Apr 27.

ISACGH: a web-based environment for the analysis of Array CGH and gene expression

which includes functional profiling.

Conde L(1), Montaner D, Burguet-Castell J, Tárraga J, Medina I, Al-Shahrour F,

Dopazo J.

Author information:

(1)Bioinformatics Department, Centro de Investigación Príncipe Felipe, INB, CIPF,

Valencia 46013, Spain.

We present the ISACGH, a web-based system that allows for the combination of

genomic data with gene expression values and provides different options for

functional profiling of the regions found. Several visualization options offer a

convenient representation of the results. Different efficient methods for

accurate estimation of genomic copy number from array-CGH hybridization data have

been included in the program. Moreover, the connection to the gene expression

analysis package GEPAS allows the use of different facilities for data

pre-processing and analysis. A DAS server allows exporting the results to the

Ensembl viewer where contextual genomic information can be obtained. The program

is freely available at: http://isacgh.bioinfo.cipf.es or within

http://www.gepas.org.

DOI: 10.1093/nar/gkm257

PMCID: PMC1933149

PMID: 17468499 [Indexed for MEDLINE]

2713. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W375-83. Epub 2007 Apr 22.

MolProbity: all-atom contacts and structure validation for proteins and nucleic

acids.

Davis IW(1), Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW,

Arendall WB 3rd, Snoeyink J, Richardson JS, Richardson DC.

Author information:

(1)Department of Biochemistry, Duke University, Durham, NC, USA.

MolProbity is a general-purpose web server offering quality validation for 3D

structures of proteins, nucleic acids and complexes. It provides detailed

all-atom contact analysis of any steric problems within the molecules as well as

updated dihedral-angle diagnostics, and it can calculate and display the H-bond

and van der Waals contacts in the interfaces between components. An integral step

in the process is the addition and full optimization of all hydrogen atoms, both

polar and nonpolar. New analysis functions have been added for RNA, for

interfaces, and for NMR ensembles. Additionally, both the web site and major

component programs have been rewritten to improve speed, convenience, clarity and

integration with other resources. MolProbity results are reported in multiple

forms: as overall numeric scores, as lists or charts of local problems, as

downloadable PDB and graphics files, and most notably as informative, manipulable

3D kinemage graphics shown online in the KiNG viewer. This service is available

free to all users at http://molprobity.biochem.duke.edu.

DOI: 10.1093/nar/gkm216

PMCID: PMC1933162

PMID: 17452350 [Indexed for MEDLINE]

2714. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W310-3. Epub 2007 Apr 22.

INFO-RNA--a server for fast inverse RNA folding satisfying sequence constraints.

Busch A(1), Backofen R.

Author information:

(1)Albert-Ludwigs-University Freiburg, Institute of Computer Science,

Bioinformatics Group, Georges-Koehler-Allee 106, 79110 Freiburg, Germany.

abusch@informatik.uni-freiburg.de

INFO-RNA is a new web server for designing RNA sequences that fold into a user

given secondary structure. Furthermore, constraints on the sequence can be

specified, e.g. one can restrict sequence positions to a fixed nucleotide or to a

set of nucleotides. Moreover, the user can allow violations of the constraints at

some positions, which can be advantageous in complicated cases. The INFO-RNA web

server allows biologists to design RNA sequences in an automatic manner. It is

clearly and intuitively arranged and easy to use. The procedure is fast, as most

applications are completed within seconds and it proceeds better and faster than

other existing tools. The INFO-RNA web server is freely available at

http://www.bioinf.uni-freiburg.de/Software/INFO-RNA/

DOI: 10.1093/nar/gkm218

PMCID: PMC1933236

PMID: 17452349 [Indexed for MEDLINE]

2715. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W12-5. Epub 2007 Apr 22.

eTBLAST: a web server to identify expert reviewers, appropriate journals and

similar publications.

Errami M(1), Wren JD, Hicks JM, Garner HR.

Author information:

(1)McDermott Center for Human Growth and Development and the Department for

Translational Research, The University of Texas Southwestern Medical Center, 5323

Harry Hines Blvd, Dallas, TX 75390-9185, USA.

Authors, editors and reviewers alike use the biomedical literature to identify

appropriate journals in which to publish, potential reviewers for papers or

grants, and collaborators (or competitors) with similar interests. Traditionally,

this process has either relied upon personal expertise and knowledge or upon a

somewhat unsystematic and laborious process of manually searching through the

literature for trends. To help with these tasks, we report three utilities that

parse and summarize the results of an abstract similarity search to find

appropriate journals for publication, authors with expertise in a given field,

and documents similar to a submitted query. The utilities are based upon a

program, eTBLAST, designed to identify similar documents within literature

databases such as (but not limited to) MEDLINE. These services are freely

accessible through the Internet at

http://invention.swmed.edu/etblast/etblast.shtml, where users can upload a file

or paste text such as an abstract into the browser interface.

DOI: 10.1093/nar/gkm221

PMCID: PMC1933238

PMID: 17452348 [Indexed for MEDLINE]

2716. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W38-42. Epub 2007 Apr 22.

Phylemon: a suite of web tools for molecular evolution, phylogenetics and

phylogenomics.

Tárraga J(1), Medina I, Arbiza L, Huerta-Cepas J, Gabaldón T, Dopazo J, Dopazo H.

Author information:

(1)Bioinformatics Department, Centro de Investigación Príncipe Felipe, INB, CIPF,

Valencia 46013, Spain.

Phylemon is an online platform for phylogenetic and evolutionary analyses of

molecular sequence data. It has been developed as a web server that integrates a

suite of different tools selected among the most popular stand-alone programs in

phylogenetic and evolutionary analysis. It has been conceived as a natural

response to the increasing demand of data analysis of many experimental

scientists wishing to add a molecular evolution and phylogenetics insight into

their research. Tools included in Phylemon cover a wide yet selected range of

programs: from the most basic for multiple sequence alignment to elaborate

statistical methods of phylogenetic reconstruction including methods for

evolutionary rates analyses and molecular adaptation. Phylemon has several

features that differentiates it from other resources: (i) It offers an integrated

environment that enables the direct concatenation of evolutionary analyses, the

storage of results and handles required data format conversions, (ii) Once an

outfile is produced, Phylemon suggests the next possible analyses, thus guiding

the user and facilitating the integration of multi-step analyses, and (iii) users

can define and save complete pipelines for specific phylogenetic analysis to be

automatically used on many genes in subsequent sessions or multiple genes in a

single session (phylogenomics). The Phylemon web server is available at

http://phylemon.bioinfo.cipf.es.

DOI: 10.1093/nar/gkm224

PMCID: PMC1933211

PMID: 17452346 [Indexed for MEDLINE]

2717. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W649-52. Epub 2007 Apr 22.

PROMALS web server for accurate multiple protein sequence alignments.

Pei J(1), Kim BH, Tang M, Grishin NV.

Author information:

(1)Howard Hughes Medical Institute, University of Texas Southwestern Medical

Center, 6001 Forest Park Road, Dallas, Texas 75390-9050, USA. jpei@chop.swmed.edu

Multiple sequence alignments are essential in homology inference, structure

modeling, functional prediction and phylogenetic analysis. We developed a web

server that constructs multiple protein sequence alignments using PROMALS, a

progressive method that improves alignment quality by using additional homologs

from PSI-BLAST searches and secondary structure predictions from PSIPRED. PROMALS

shows higher alignment accuracy than other advanced methods, such as MUMMALS,

ProbCons, MAFFT and SPEM. The PROMALS web server takes FASTA format protein

sequences as input. The output includes a colored alignment augmented with

information about sequence grouping, predicted secondary structures and

positional conservation. The PROMALS web server is available at:

http://prodata.swmed.edu/promals/

DOI: 10.1093/nar/gkm227

PMCID: PMC1933189

PMID: 17452345 [Indexed for MEDLINE]

2718. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W126-31. Epub 2007 Apr 16.

OPTIMIZER: a web server for optimizing the codon usage of DNA sequences.

Puigbò P(1), Guzmán E, Romeu A, Garcia-Vallvé S.

Author information:

(1)Evolutionary Genomics Group, Biochemistry and Biotechnology Department,

Faculty of Chemistry, Rovira i Virgili University (URV), Tarragona, Spain.

OPTIMIZER is an on-line application that optimizes the codon usage of a gene to

increase its expression level. Three methods of optimization are available: the

'one amino acid-one codon' method, a guided random method based on a Monte Carlo

algorithm, and a new method designed to maximize the optimization with the fewest

changes in the query sequence. One of the main features of OPTIMIZER is that it

makes it possible to optimize a DNA sequence using pre-computed codon usage

tables from a predicted group of highly expressed genes from more than 150

prokaryotic species under strong translational selection. These groups of highly

expressed genes have been predicted using a new iterative algorithm. In addition,

users can use, as a reference set, a pre-computed table containing the mean codon

usage of ribosomal protein genes and, as a novelty, the tRNA gene-copy numbers.

OPTIMIZER is accessible free of charge at http://genomes.urv.es/OPTIMIZER.

DOI: 10.1093/nar/gkm219

PMCID: PMC1933141

PMID: 17439967 [Indexed for MEDLINE]

2719. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W148-51. Epub 2007 Apr 16.

Wheat Estimated Transcript Server (WhETS): a tool to provide best estimate of

hexaploid wheat transcript sequence.

Mitchell RA(1), Castells-Brooke N, Taubert J, Verrier PJ, Leader DJ, Rawlings CJ.

Author information:

(1)Biomathematics and Bioinformatics Division, Rothamsted Research, Harpenden,

Hertfordshire AL5 2JQ, UK. rowan.mitchell@bbsrc.ac.uk

Wheat biologists face particular problems because of the lack of genomic sequence

and the three homoeologous genomes which give rise to three very similar forms

for many transcripts. However, over 1.3 million available public-domain Triticeae

ESTs (of which approximately 850,000 are wheat) and the full rice genomic

sequence can be used to estimate likely transcript sequences present in any wheat

cDNA sample to which PCR primers may then be designed. Wheat Estimated Transcript

Server (WhETS) is designed to do this in a convenient form, and to provide

information on the number of matching EST and high quality cDNA (hq-cDNA)

sequences, tissue distribution and likely intron position inferred from rice.

Triticeae EST and hq-cDNA sequences are mapped onto rice loci and stored in a

database. The user selects a rice locus (directly or via Arabidopsis) and the

matching Triticeae sequences are assembled according to user-defined filter and

stringency settings. Assembly is achieved initially with the CAP3 program and

then with a single nucleotide polymorphism (SNP)-analysis algorithm designed to

separate homoeologues. Alignment of the resulting contigs and singlets against

the rice template sequence is then displayed. Sequences and assembly details are

available for download in fasta and ace formats, respectively. WhETS is

accessible at http://www4.rothamsted.bbsrc.ac.uk/whets.

DOI: 10.1093/nar/gkm220

PMCID: PMC1933201

PMID: 17439966 [Indexed for MEDLINE]

2720. PLoS Comput Biol. 2007 Jul;3(7):e119.

Protein-protein interaction hotspots carved into sequences.

Ofran Y(1), Rost B.

Author information:

(1)Department of Biochemistry and Molecular Biophysics, Columbia University, New

York, New York, USA. yo135@columbia.edu

Protein-protein interactions, a key to almost any biological process, are

mediated by molecular mechanisms that are not entirely clear. The study of these

mechanisms often focuses on all residues at protein-protein interfaces. However,

only a small subset of all interface residues is actually essential for

recognition or binding. Commonly referred to as "hotspots," these essential

residues are defined as residues that impede protein-protein interactions if

mutated. While no in silico tool identifies hotspots in unbound chains, numerous

prediction methods were designed to identify all the residues in a protein that

are likely to be a part of protein-protein interfaces. These methods typically

identify successfully only a small fraction of all interface residues. Here, we

analyzed the hypothesis that the two subsets correspond (i.e., that in silico

methods may predict few residues because they preferentially predict hotspots).

We demonstrate that this is indeed the case and that we can therefore predict

directly from the sequence of a single protein which residues are interaction

hotspots (without knowledge of the interaction partner). Our results suggested

that most protein complexes are stabilized by similar basic principles. The

ability to accurately and efficiently identify hotspots from sequence enables the

annotation and analysis of protein-protein interaction hotspots in entire

organisms and thus may benefit function prediction and drug development. The

server for prediction is available at http://www.rostlab.org/services/isis.

DOI: 10.1371/journal.pcbi.0030119

PMCID: PMC1914369

PMID: 17630824 [Indexed for MEDLINE]

2721. Proteins. 2007 Jul 1;68(1):76-81.

Real-SPINE: an integrated system of neural networks for real-value prediction of

protein structural properties.

Dor O(1), Zhou Y.

Author information:

(1)Department of Physiology and Biophysics, Howard Hughes Medical Institute

Center for Single Molecule Biophysics, State University of New York at Buffalo,

Buffalo, New York 14214, USA.

Proteins can move freely in three-dimensional space. As a result, their

structural properties, such as solvent accessible surface area, backbone dihedral

angles, and atomic distances, are continuous variables. However, these properties

are often arbitrarily divided into a few classes to facilitate prediction by

statistical learning techniques. In this work, we establish an integrated system

of neural networks (called Real-SPINE) for real-value prediction and apply the

method to predict residue-solvent accessibility and backbone psi dihedral angles

of proteins based on information derived from sequences only. Real-SPINE is

trained with a large data set of 2640 protein chains, sequence profiles generated

from multiple sequence alignment, representative amino-acid properties, a slow

learning rate, overfitting protection, and predicted secondary structures. The

method optimizes more than 200,000 weights and yields a 10-fold cross-validated

Pearson's correlation coefficient (PCC) of 0.74 between predicted and actual

solvent accessible surface areas and 0.62 between predicted and actual psi

angles. In particular, 90% of 2640 proteins have a PCC value greater than 0.6

between predicted and actual solvent-accessible surface areas. The results of

Real-SPINE can be compared with the best reported correlation coefficients of

0.64-0.67 for solvent-accessible surface areas and 0.47 for psi angles. The

real-SPINE server, executable programs, and datasets are freely available on

http://sparks.informatics.iupui.edu.

2007 Wiley-Liss, Inc.

DOI: 10.1002/prot.21408

PMID: 17397056 [Indexed for MEDLINE]

2722. Bioinformatics. 2007 Jun 15;23(12):1559-61. Epub 2007 Mar 30.

SCEPTRANS: an online tool for analyzing periodic transcription in yeast.

Kudlicki A(1), Rowicka M, Otwinowski Z.

Author information:

(1)University of Texas Southwestern Medical Center, Department of Biochemistry,

Dallas, TX 75390-8816, USA.

SUMMARY: SCEPTRANS is designed for analysis of microarray timecourse data related

to periodic phenomena in the budding yeast. The server allows for easy viewing of

temporal profiles of multiple genes in a number of datasets. Additional

functionality includes searching for coexpressed genes, periodicity and

correlation analysis, integrating functional annotation and localization data as

well as advanced operations on sets of genes.

AVAILABILITY: Available online at http://sceptrans.org/

DOI: 10.1093/bioinformatics/btm126

PMID: 17400726 [Indexed for MEDLINE]

2723. Bioinformatics. 2007 Jun 15;23(12):1444-50. Epub 2007 Mar 24.

Finding new structural and sequence attributes to predict possible disease

association of single amino acid polymorphism (SAP).

Ye ZQ(1), Zhao SQ, Gao G, Liu XQ, Langlois RE, Lu H, Wei L.

Author information:

(1)Center for Bioinformatics, National Laboratory of Protein Engineering and

Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing,

PR China.

MOTIVATION: The rapid accumulation of single amino acid polymorphisms (SAPs),

also known as non-synonymous single nucleotide polymorphisms (nsSNPs), brings the

opportunities and needs to understand and predict their disease association.

Currently published attributes are limited, the detailed mechanisms governing the

disease association of a SAP remain unclear and thus, further investigation of

new attributes and improvement of the prediction are desired.

RESULTS: A SAP dataset was compiled from the Swiss-Prot variant pages. We

extracted and demonstrated the effectiveness of several new biologically

informative attributes including the structural neighbor profiles that describe

the SAP's microenvironment, nearby functional sites that measure the

structure-based and sequence-based distances between the SAP site and its nearby

functional sites, aggregation properties that measure the likelihood of protein

aggregation and disordered regions that consider whether the SAP is located in

structurally disordered regions. The new attributes provided insights into the

mechanisms of the disease association of SAPs. We built a support vector machines

(SVMs) classifier employing a carefully selected set of new and previously

published attributes. Through a strict protein-level 5-fold cross-validation, we

attained an overall accuracy of 82.61%, and an MCC of 0.60. Moreover, a web

server was developed to provide a user-friendly interface for biologists.

AVAILABILITY: The web server is available at http://sapred.cbi.pku.edu.cn/

DOI: 10.1093/bioinformatics/btm119

PMID: 17384424 [Indexed for MEDLINE]

2724. Bioinformatics. 2007 Jun 15;23(12):1568-70. Epub 2007 Jan 18.

ProServer: a simple, extensible Perl DAS server.

Finn RD(1), Stalker JW, Jackson DK, Kulesha E, Clements J, Pettett R.

Author information:

(1)Wellcome Trust Sanger Institute, Wellcome Trust Geome Campus, Hinxton,

Cambridge, UK.

SUMMARY: The increasing size and complexity of biological databases has led to a

growing trend to federate rather than duplicate them. In order to share data

between federated databases, protocols for the exchange mechanism must be

developed. One such data exchange protocol that is widely used is the Distributed

Annotation System (DAS). For example, DAS has enabled small experimental groups

to integrate their data into the Ensembl genome browser. We have developed

ProServer, a simple, lightweight, Perl-based DAS server that does not depend on a

separate HTTP server. The ProServer package is easily extensible, allowing data

to be served from almost any underlying data model. Recent additions to the DAS

protocol have enabled both structure and alignment (sequence and structural) data

to be exchanged. ProServer allows both of these data types to be served.

AVAILABILITY: ProServer can be downloaded from http://www.sanger.ac.uk/proserver/

or CPAN http://search.cpan.org/~rpettett/. Details on the system requirements and

installation of ProServer can be found at http://www.sanger.ac.uk/proserver/.

DOI: 10.1093/bioinformatics/btl650

PMCID: PMC2989875

PMID: 17237073 [Indexed for MEDLINE]

2725. Biochem Biophys Res Commun. 2007 Jun 8;357(3):633-40. Epub 2007 Apr 5.

Signal-CF: a subsite-coupled and window-fusing approach for predicting signal

peptides.

Chou KC(1), Shen HB.

Author information:

(1)Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA

92130, USA. kchou@san.rr.com

We have developed an automated method for predicting signal peptide sequences and

their cleavage sites in eukaryotic and bacterial protein sequences. It is a

2-layer predictor: the 1st-layer prediction engine is to identify a query protein

as secretory or non-secretory; if it is secretory, the process will be

automatically continued with the 2nd-layer prediction engine to further identify

the cleavage site of its signal peptide. The new predictor is called Signal-CF,

where C stands for "coupling" and F for "fusion", meaning that Signal-CF is

formed by incorporating the subsite coupling effects along a protein sequence and

by fusing the results derived from many width-different scaled windows through a

voting system. Signal-CF is featured by high success prediction rates with short

computational time, and hence is particularly useful for the analysis of

large-scale datasets. Signal-CF is freely available as a web-server at

http://chou.med.harvard.edu/bioinf/Signal-CF/ or

http://202.120.37.186/bioinf/Signal-CF/.

DOI: 10.1016/j.bbrc.2007.03.162

PMID: 17434148 [Indexed for MEDLINE]

2726. BMC Bioinformatics. 2007 Jun 8;8:189.

MATLIGN: a motif clustering, comparison and matching tool.

Kankainen M(1), Löytynoja A.

Author information:

(1)Institute of Biotechnology, University of Helsinki, Helsinki, Finland.

matti.kankainen@helsinki.fi

BACKGROUND: Sequence motifs representing transcription factor binding sites

(TFBS) are commonly encoded as position frequency matrices (PFM) or degenerate

consensus sequences (CS). These formats are used to represent the characterised

TFBS profiles stored in transcription factor databases, as well as to represent

the potential motifs predicted using computational methods. To fill the gap

between the known and predicted motifs, methods are needed for the

post-processing of prediction results, i.e. for matching, comparison and

clustering of pre-selected motifs. The computational identification of

over-represented motifs in sets of DNA sequences is, in particular, a task where

post-processing can dramatically simplify the analysis. Efficient

post-processing, for example, reduces the redundancy of the motifs predicted and

enables them to be annotated.

RESULTS: In order to facilitate the post-processing of motifs, in both PFM and CS

formats, we have developed a tool called Matlign. The tool aligns and evaluates

the similarity of motifs using a combination of scoring functions, and visualises

the results using hierarchical clustering. By limiting the number of distinct

gaps created (though, not their length), the alignment algorithm also correctly

aligns motifs with an internal spacer. The method selects the best non-redundant

motif set, with repetitive motifs merged together, by cutting the hierarchical

tree using silhouette values. Our analyses show that Matlign can reliably

discover the most similar analogue from a collection of characterised regulatory

elements such that the method is also useful for the annotation of motif

predictions by PFM library searches.

CONCLUSION: Matlign is a user-friendly tool for post-processing large collections

of DNA sequence motifs. Starting from a large number of potential regulatory

motifs, Matlign provides a researcher with a non-redundant set of motifs, which

can then be further associated to known regulatory elements. A web-server is

available at http://ekhidna.biocenter.helsinki.fi/poxo/matlign.

DOI: 10.1186/1471-2105-8-189

PMCID: PMC1925120

PMID: 17559640 [Indexed for MEDLINE]

2727. Biotechnol Bioeng. 2007 Jun 1;97(2):389-96.

MODEL-molecular descriptor lab: a web-based server for computing structural and

physicochemical features of compounds.

Li ZR(1), Han LY, Xue Y, Yap CW, Li H, Jiang L, Chen YZ.

Author information:

(1)Bioinformatics and Drug Design Group, Department of Computational Science,

National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2,

Singapore.

Molecular descriptors represent structural and physicochemical features of

compounds. They have been extensively used for developing statistical models,

such as quantitative structure activity relationship (QSAR) and artificial neural

networks (NN), for computer prediction of the pharmacodynamic, pharmacokinetic,

or toxicological properties of compounds from their structure. While computer

programs have been developed for computing molecular descriptors, there is a lack

of a freely accessible one. We have developed a web-based server, MODEL

(Molecular Descriptor Lab), for computing a comprehensive set of 3,778 molecular

descriptors, which is significantly more than the approximately 1,600 molecular

descriptors computed by other software. Our computational algorithms have been

extensively tested and the computed molecular descriptors have been used in a

number of published works of statistical models for predicting variety of

pharmacodynamic, pharmacokinetic, and toxicological properties of compounds.

Several testing studies on the computed molecular descriptors are discussed.

MODEL is accessible at http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi free of

charge for academic use.

(c) 2006 Wiley Periodicals, Inc.

DOI: 10.1002/bit.21214

PMID: 17013940 [Indexed for MEDLINE]

2728. J Comput Biol. 2007 Jun;14(5):637-54.

A reduction-based exact algorithm for the contact map overlap problem.

Xie W(1), Sahinidis NV.

Author information:

(1)Department of Chemical and Biomolecular Engineering, University of Illinois at

Urbana-Champaign, Urbana, Illinois 61801, USA.

Aligning proteins based on their structural similarity is a fundamental problem

in molecular biology with applications in many settings, including structure

classification, database search, function prediction, and assessment of folding

prediction methods. Structural alignment can be done via several methods,

including contact map overlap (CMO) maximization that aligns proteins in a way

that maximizes the number of common residue contacts. In this paper, we develop a

reduction-based exact algorithm for the CMO problem. Our approach solves CMO

directly rather than after transformation to other combinatorial optimization

problems. We exploit the mathematical structure of the problem in order to

develop a number of efficient lower bounding, upper bounding, and reduction

schemes. Computational experiments demonstrate that our algorithm runs

significantly faster than existing exact algorithms and solves some hard CMO

instances that were not solved in the past. In addition, the algorithm produces

protein clusters that are in excellent agreement with the SCOP classification. An

implementation of our algorithm is accessible as an on-line server at

http://eudoxus.scs.uiuc.edu/cmos/cmos.html.

DOI: 10.1089/cmb.2007.R007

PMID: 17683265 [Indexed for MEDLINE]

2729. Phytochemistry. 2007 Jun;68(12):1605-11. Epub 2007 Jun 4.

PeroxiBase: the peroxidase database.

Passardi F(1), Theiler G, Zamocky M, Cosio C, Rouhier N, Teixera F,

Margis-Pinheiro M, Ioannidis V, Penel C, Falquet L, Dunand C.

Author information:

(1)Laboratory of Plant Physiology, University of Geneva, Quai Ernest-Ansermet 30,

CH-1211 Geneva 4, Switzerland.

Peroxidases (EC 1.11.1.x), which are encoded by small or large multigenic

families, are involved in several important physiological and developmental

processes. Analyzing their evolution and their distribution among various phyla

could certainly help to elucidate the mystery of their extremely widespread and

diversified presence in almost all living organisms. PeroxiBase was originally

created for the exhaustive collection of class III peroxidase sequences from

plants (Bakalovic, N., Passardi, F., et al., 2006. PeroxiBase: a class III plant

peroxidase database. Phytochemistry 67, 534-539). The extension of the class III

peroxidase database to all proteins capable to reduce peroxide molecules appears

as a necessity. Our database contains haem and non-haem peroxidase sequences

originated from annotated or not correctly annotated sequences deposited in the

main repositories such as GenBank or UniProt KnowledgeBase. This new database

will allow obtaining a global overview of the evolution the protein families and

superfamilies capable of peroxidase reaction. In this rapidly growing field,

there is a need for continual updates and corrections of the peroxidase protein

sequences. Following the lack of unified nomenclature, we also introduced a

unique abbreviation for each different family of peroxidases. This paper thus

aims to report the evolution of the PeroxiBase database, which is freely

accessible through a web server (http://peroxibase.isb-sib.ch). In addition to

new categories of peroxidases, new specific tools have been created to facilitate

query, classification and submission of peroxidase sequences.

DOI: 10.1016/j.phytochem.2007.04.005

PMID: 17544465 [Indexed for MEDLINE]

2730. Proteins. 2007 Jun 1;67(4):1167-78.

Computational protocol for predicting the binding affinities of zinc containing

metalloprotein-ligand complexes.

Jain T(1), Jayaram B.

Author information:

(1)Department of Chemistry, Indian Institute of Technology, Hauz Khas, New

Delhi-110016, India.

Zinc is one of the most important metal ions found in proteins performing

specific functions associated with life processes. Coordination geometry of the

zinc ion in the active site of the metalloprotein-ligand complexes poses a

challenge in determining ligand binding affinities accurately in structure-based

drug design. We report here an all atom force field based computational protocol

for estimating rapidly the binding affinities of zinc containing

metalloprotein-ligand complexes, considering electrostatics, van der Waals,

hydrophobicity, and loss in conformational entropy of protein side chains upon

ligand binding along with a nonbonded approach to model the interactions of the

zinc ion with all the other atoms of the complex. We examined the sensitivity of

the binding affinity predictions to the choice of Lennard-Jones parameters,

partial atomic charges, and dielectric treatments adopted for system preparation

and scoring. The highest correlation obtained was R2 = 0.77 (r = 0.88) for the

predicted binding affinity against the experiment on a heterogenous dataset of 90

zinc containing metalloprotein-ligand complexes consisting of five unique protein

targets. Model validation and parameter analysis studies underscore the

robustness and predictive ability of the scoring function. The high correlation

obtained suggests the potential applicability of the methodology in designing

novel ligands for zinc-metalloproteins. The scoring function has been web enabled

for free access at www.scfbio-iitd.res.in/software/drugdesign/bapplz.jsp as

BAPPL-Z server (Binding Affinity Prediction of Protein-Ligand complexes

containing Zinc metal ions).

2007 Wiley-Liss, Inc.

DOI: 10.1002/prot.21332

PMID: 17380508 [Indexed for MEDLINE]

2731. BMC Bioinformatics. 2007 May 22;8 Suppl 4:S3.

CORRIE: enzyme sequence annotation with confidence estimates.

Audit B(1), Levy ED, Gilks WR, Goldovsky L, Ouzounis CA.

Author information:

(1)Laboratoire Joliot-Curie and Laboratoire de Physique, CNRS UMR5672, Ecole

Normale Supérieure, Lyon CEDEX 07, France. benjamin.audit@ens-lyon.fr

Using a previously developed automated method for enzyme annotation, we report

the re-annotation of the ENZYME database and the analysis of local error rates

per class. In control experiments, we demonstrate that the method is able to

correctly re-annotate 91% of all Enzyme Classification (EC) classes with high

coverage (755 out of 827). Only 44 enzyme classes are found to contain false

positives, while the remaining 28 enzyme classes are not represented. We also

show cases where the re-annotation procedure results in partial overlaps for

those few enzyme classes where a certain inconsistency might appear between

homologous proteins, mostly due to function specificity. Our results allow the

interactive exploration of the EC hierarchy for known enzyme families as well as

putative enzyme sequences that may need to be classified within the EC hierarchy.

These aspects of our framework have been incorporated into a web-server, called

CORRIE, which stands for Correspondence Indicator Estimation and allows the

interactive prediction of a functional class for putative enzymes from sequence

alone, supported by probabilistic measures in the context of the pre-calculated

Correspondence Indicators of known enzymes with the functional classes of the EC

hierarchy. The CORRIE server is available at:

http://www.genomes.org/services/corrie/.

DOI: 10.1186/1471-2105-8-S4-S3

PMCID: PMC1892082

PMID: 17570146 [Indexed for MEDLINE]

2732. BMC Bioinformatics. 2007 May 22;8 Suppl 4:S2.

SVM-Fold: a tool for discriminative multi-class protein fold and superfamily

recognition.

Melvin I(1), Ie E, Kuang R, Weston J, Stafford WN, Leslie C.

Author information:

(1)NEC Laboratories of America, Princeton, NJ, USA. imelvin@gmail.com

BACKGROUND: Predicting a protein's structural class from its amino acid sequence

is a fundamental problem in computational biology. Much recent work has focused

on developing new representations for protein sequences, called string kernels,

for use with support vector machine (SVM) classifiers. However, while some of

these approaches exhibit state-of-the-art performance at the binary protein

classification problem, i.e. discriminating between a particular protein class

and all other classes, few of these studies have addressed the real problem of

multi-class superfamily or fold recognition. Moreover, there are only limited

software tools and systems for SVM-based protein classification available to the

bioinformatics community.

RESULTS: We present a new multi-class SVM-based protein fold and superfamily

recognition system and web server called SVM-Fold, which can be found at

http://svm-fold.c2b2.columbia.edu. Our system uses an efficient implementation of

a state-of-the-art string kernel for sequence profiles, called the profile

kernel, where the underlying feature representation is a histogram of inexact

matching k-mer frequencies. We also employ a novel machine learning approach to

solve the difficult multi-class problem of classifying a sequence of amino acids

into one of many known protein structural classes. Binary one-vs-the-rest SVM

classifiers that are trained to recognize individual structural classes yield

prediction scores that are not comparable, so that standard "one-vs-all"

classification fails to perform well. Moreover, SVMs for classes at different

levels of the protein structural hierarchy may make useful predictions, but

one-vs-all does not try to combine these multiple predictions. To deal with these

problems, our method learns relative weights between one-vs-the-rest classifiers

and encodes information about the protein structural hierarchy for multi-class

prediction. In large-scale benchmark results based on the SCOP database, our code

weighting approach significantly improves on the standard one-vs-all method for

both the superfamily and fold prediction in the remote homology setting and on

the fold recognition problem. Moreover, our code weight learning algorithm

strongly outperforms nearest-neighbor methods based on PSI-BLAST in terms of

prediction accuracy on every structure classification problem we consider.

CONCLUSION: By combining state-of-the-art SVM kernel methods with a novel

multi-class algorithm, the SVM-Fold system delivers efficient and accurate

protein fold and superfamily recognition.

DOI: 10.1186/1471-2105-8-S4-S2

PMCID: PMC1892081

PMID: 17570145 [Indexed for MEDLINE]

2733. BMC Bioinformatics. 2007 May 21;8:160.

TISs-ST: a web server to evaluate polymorphic translation initiation sites and

their reflections on the secretory targets.

Vicentini R(1), Menossi M.

Author information:

(1)Functional Genomics Laboratory, Center for Molecular Biology and Genetic

Engineering, State University of Campinas, Campinas, SP, Brazil.

shinapes@unicamp.br <shinapes@unicamp.br>

BACKGROUND: The nucleotide sequence flanking the translation initiation codon

(start codon context) affects the translational efficiency of eukaryotic mRNAs,

and may indicate the presence of an alternative translation initiation site (TIS)

to produce proteins with different properties. Multi-targeting may reflect the

translational variability of these other protein forms. In this paper we present

a web server that performs computations to investigate the usage of alternative

translation initiation sites for the synthesis of new protein variants that might

have different functions.

RESULTS: An efficient web-based tool entitled TISs-ST (Translation Initiation

Sites and Secretory Targets) evaluates putative translation initiation sites and

indicates the prediction of a signal peptide of the protein encoded from this

site. The TISs-ST web server is freely available to both academic and commercial

users and can be accessed at http://ipe.cbmeg.unicamp.br/pub/TISs-ST.

CONCLUSION: The program can be used to evaluate alternative translation

initiation site consensus with user-specified sequences, based on their

composition or on many position weight matrix models. TISs-ST provides analytical

and visualization tools for evaluating the periodic frequency, the consensus

pattern and the total information content of a sequence data set. A search option

allows for the identification of signal peptides from predicted proteins using

the PrediSi software.

DOI: 10.1186/1471-2105-8-160

PMCID: PMC1891115

PMID: 17517132 [Indexed for MEDLINE]

2734. Bioinformatics. 2007 May 15;23(10):1195-202. Epub 2007 Mar 28.

A fast and flexible approach to oligonucleotide probe design for genomes and gene

families.

Feng S(1), Tillier ER.

Author information:

(1)Institute of Computing Technology, Chinese Academy of Sciences, China.

MOTIVATION: With hundreds of completely sequenced microbial genomes available,

and advancements in DNA microarray technology, the detection of genes in

microbial communities consisting of hundreds of thousands of sequences may be

possible. The existing strategies developed for DNA probe design, geared toward

identifying specific sequences, are not suitable due to the lack of coverage,

flexibility and efficiency necessary for applications in metagenomics.

METHODS: ProDesign is a tool developed for the selection of oligonucleotide

probes to detect members of gene families present in environmental samples. Gene

family-specific probe sequences are generated based on specific and shared words,

which are found with the spaced seed hashing algorithm. To detect more sequences,

those sharing some common words are re-clustered into new families, then probes

specific for the new families are generated.

RESULTS: The program is very flexible in that it can be used for designing probes

for detecting many genes families simultaneously and specifically in one or more

genomes. Neither the length nor the melting temperature of the probes needs to be

predefined. We have found that ProDesign provides more flexibility, coverage and

speed than other software programs used in the selection of probes for genomic

and gene family arrays.

AVAILABILITY: ProDesign is licensed free of charge to academic users. ProDesign

and Supplementary Material can be obtained by contacting the authors. A web

server for ProDesign is available at

http://www.uhnresearch.ca/labs/tillier/ProDesign/ProDesign.html.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btm114

PMID: 17392329 [Indexed for MEDLINE]

2735. Bioinformatics. 2007 May 15;23(10):1203-10. Epub 2007 Mar 22.

AutoSCOP: automated prediction of SCOP classifications using unique pattern-class

mappings.

Gewehr JE(1), Hintermair V, Zimmer R.

Author information:

(1)Practical Informatics and Bioinformatics Group, Department of Informatics,

Ludwig-Maximilians-University Munich, Amalienstr. 17, D-80333 Munich, Germany.

jan.gewehr@ifi.lmu.de

MOTIVATION: The sequence patterns contained in the available motif and hidden

Markov model (HMM) databases are a valuable source of information for protein

sequence annotation. For structure prediction and fold recognition purposes, we

computed mappings from such pattern databases to the protein domain hierarchy

given by the ASTRAL compendium and applied them to the prediction of SCOP

classifications. Our aim is to make highly confident predictions also for

non-trivial cases if possible and abstain from a prediction otherwise, and thus

to provide a method that can be used as a first step in a pipeline of prediction

methods. We describe two successful examples for such pipelines. With the

AutoSCOP approach, it is possible to make predictions in a large-scale manner for

many domains of the available sequences in the well-known protein sequence

databases.

RESULTS: AutoSCOP computes unique sequence patterns and pattern combinations for

SCOP classifications. For instance, we assign a SCOP superfamily to a pattern

found in its members whenever the pattern does not occur in any other SCOP

superfamily. Especially on the fold and superfamily level, our method achieves

both high sensitivity (above 93%) and high specificity (above 98%) on the

difference set between two ASTRAL versions, due to being able to abstain from

unreliable predictions. Further, on a harder test set filtered at low sequence

identity, the combination with profile-profile alignments improves accuracy and

performs comparably even to structure alignment methods. Integrating our method

with structure alignment, we are able to achieve an accuracy of 99% on SCOP fold

classifications on this set. In an analysis of false assignments of domains from

new folds/superfamilies/families to existing SCOP classifications, AutoSCOP

correctly abstains for more than 70% of the domains belonging to new folds and

superfamilies, and more than 80% of the domains belonging to new families. These

findings show that our approach is a useful additional filter for SCOP

classification prediction of protein domains in combination with well-known

methods such as profile-profile alignment.

AVAILABILITY: A web server where users can input their domain sequences is

available at http://www.bio.ifi.lmu.de/autoscop.

DOI: 10.1093/bioinformatics/btm089

PMID: 17379694 [Indexed for MEDLINE]

2736. Bioinformatics. 2007 May 15;23(10):1181-7. Epub 2007 Mar 22.

IMEx: Imperfect Microsatellite Extractor.

Mudunuri SB(1), Nagarajaram HA.

Author information:

(1)Laboratory of Computational Biology, Centre for DNA Fingerprinting and

Diagnostics, ECIL Road, Nacharam, Hyderabad 500 076, India.

MOTIVATION: Microsatellites, also known as simple sequence repeats, are the

tandem repeats of nucleotide motifs of the size 1-6 bp found in every genome

known so far. Their importance in genomes is well known. Microsatellites are

associated with various disease genes, have been used as molecular markers in

linkage analysis and DNA fingerprinting studies, and also seem to play an

important role in the genome evolution. Therefore, it is of importance to study

distribution, enrichment and polymorphism of microsatellites in the genomes of

interest. For this, the prerequisite is the availability of a computational tool

for extraction of microsatellites (perfect as well as imperfect) and their

related information from whole genome sequences. Examination of available tools

revealed certain lacunae in them and prompted us to develop a new tool.

RESULTS: In order to efficiently screen genome sequences for microsatellites

(perfect as well as imperfect), we developed a new tool called IMEx (Imperfect

Microsatellite Extractor). IMEx uses simple string-matching algorithm with

sliding window approach to screen DNA sequences for microsatellites and reports

the motif, copy number, genomic location, nearby genes, mutational events and

many other features useful for in-depth studies. IMEx is more sensitive,

efficient and useful than the available widely used tools. IMEx is available in

the form of a stand-alone program as well as in the form of a web-server.

AVAILABILITY: A World Wide Web server and the stand-alone program are available

for free access at http://203.197.254.154/IMEX/ or http://www.cdfd.org.in/imex.

DOI: 10.1093/bioinformatics/btm097

PMID: 17379689 [Indexed for MEDLINE]

2737. Bioinformatics. 2007 May 15;23(10):1292-3. Epub 2007 Mar 22.

iPTREE-STAB: interpretable decision tree based method for predicting protein

stability changes upon mutations.

Huang LT(1), Gromiha MM, Ho SY.

Author information:

(1)Department of Computer Science and Information Engineering, Ming-Dao

University, Changhua 523, Taiwan.

We have developed a web server, iPTREE-STAB for discriminating the stability of

proteins (stabilizing or destabilizing) and predicting their stability changes

(delta deltaG) upon single amino acid substitutions from amino acid sequence. The

discrimination and prediction are mainly based on decision tree coupled with

adaptive boosting algorithm, and classification and regression tree,

respectively, using three neighboring residues of the mutant site along N- and

C-terminals. Our method showed an accuracy of 82% for discriminating the

stabilizing and destabilizing mutants, and a correlation of 0.70 for predicting

protein stability changes upon mutations.AVAILABILITY:

http://bioinformatics.myweb.hinet.net/iptree.htm.

SUPPLEMENTARY INFORMATION: Dataset and other details are given.

DOI: 10.1093/bioinformatics/btm100

PMID: 17379687 [Indexed for MEDLINE]

2738. BMC Bioinformatics. 2007 May 15;8:155.

TAP score: torsion angle propensity normalization applied to local protein

structure evaluation.

Tosatto SC(1), Battistutta R.

Author information:

(1)Dept. of Biology and CRIBI Biotechnology Centre, University of Padova, Italy.

silvio.tosatto@unipd.it

BACKGROUND: Experimentally determined protein structures may contain errors and

require validation. Conformational criteria based on the Ramachandran plot are

mainly used to distinguish between distorted and adequately refined models. While

the readily available criteria are sufficient to detect totally wrong structures,

establishing the more subtle differences between plausible structures remains

more challenging.

RESULTS: A new criterion, called TAP score, measuring local sequence to structure

fitness based on torsion angle propensities normalized against the global minimum

and maximum is introduced. It is shown to be more accurate than previous methods

at estimating the validity of a protein model in terms of commonly used

experimental quality parameters on two test sets representing the full PDB

database and a subset of obsolete PDB structures. Highly selective TAP thresholds

are derived to recognize over 90% of the top experimental structures in the

absence of experimental information. Both a web server and an executable version

of the TAP score are available at http://protein.cribi.unipd.it/tap/.

CONCLUSION: A novel procedure for energy normalization (TAP) has significantly

improved the possibility to recognize the best experimental structures. It will

allow the user to more reliably isolate problematic structures in the context of

automated experimental structure determination.

DOI: 10.1186/1471-2105-8-155

PMCID: PMC1878508

PMID: 17504537 [Indexed for MEDLINE]

2739. BMC Bioinformatics. 2007 May 9;8:151.

Tracembler--software for in-silico chromosome walking in unassembled genomes.

Dong Q(1), Wilkerson MD, Brendel V.

Author information:

(1)Department of Genetics, Development & Cell Biology, Iowa State University,

Ames, Iowa 50011, USA. qfdong@cgb.indiana.edu <qfdong@cgb.indiana.edu>

BACKGROUND: Whole genome shotgun sequencing produces increasingly higher coverage

of a genome with random sequence reads. Progressive whole genome assembly and

eventual finishing sequencing is a process that typically takes several years for

large eukaryotic genomes. In the interim, all sequence reads of public sequencing

projects are made available in repositories such as the NCBI Trace Archive. For a

particular locus, sequencing coverage may be high enough early on to produce a

reliable local genome assembly. We have developed software, Tracembler, that

facilitates in silico chromosome walking by recursively assembling reads of a

selected species from the NCBI Trace Archive starting with reads that

significantly match sequence seeds supplied by the user.

RESULTS: Tracembler takes one or multiple DNA or protein sequence(s) as input to

the NCBI Trace Archive BLAST engine to identify matching sequence reads from a

species of interest. The BLAST searches are carried out recursively such that

BLAST matching sequences identified in previous rounds of searches are used as

new queries in subsequent rounds of BLAST searches. The recursive BLAST search

stops when either no more new matching sequences are found, a given maximal

number of queries is exhausted, or a specified maximum number of rounds of

recursion is reached. All the BLAST matching sequences are then assembled into

contigs based on significant sequence overlaps using the CAP3 program. We

demonstrate the validity of the concept and software implementation with an

example of successfully recovering a full-length Chrm2 gene as well as its

upstream and downstream genomic regions from Rattus norvegicus reads. In a second

example, a query with two adjacent Medicago truncatula genes as seeds resulted in

a contig that likely identifies the microsyntenic homologous soybean locus.

CONCLUSION: Tracembler streamlines the process of recursive database searches,

sequence assembly, and gene identification in resulting contigs in attempts to

identify homologous loci of genes of interest in species with emerging whole

genome shotgun reads. A web server hosting Tracembler is provided at

http://www.plantgdb.org/tool/tracembler/, and the software is also freely

available from the authors for local installations.

DOI: 10.1186/1471-2105-8-151

PMCID: PMC1876249

PMID: 17490482 [Indexed for MEDLINE]

2740. BMC Biol. 2007 May 8;5:17.

Ab initio modeling of small proteins by iterative TASSER simulations.

Wu S(1), Skolnick J, Zhang Y.

Author information:

(1)Center for Bioinformatics and Department of Molecular Bioscience, University

of Kansas, Lawrence, KS 66047, USA. stwu@ku.edu

BACKGROUND: Predicting 3-dimensional protein structures from amino-acid sequences

is an important unsolved problem in computational structural biology. The problem

becomes relatively easier if close homologous proteins have been solved, as

high-resolution models can be built by aligning target sequences to the solved

homologous structures. However, for sequences without similar folds in the

Protein Data Bank (PDB) library, the models have to be predicted from scratch.

Progress in the ab initio structure modeling is slow. The aim of this study was

to extend the TASSER (threading/assembly/refinement) method for the ab initio

modeling and examine systemically its ability to fold small single-domain

proteins.

RESULTS: We developed I-TASSER by iteratively implementing the TASSER method,

which is used in the folding test of three benchmarks of small proteins. First,

data on 16 small proteins (< 90 residues) were used to generate I-TASSER models,

which had an average Calpha-root mean square deviation (RMSD) of 3.8A, with 6 of

them having a Calpha-RMSD < 2.5A. The overall result was comparable with the

all-atomic ROSETTA simulation, but the central processing unit (CPU) time by

I-TASSER was much shorter (150 CPU days vs. 5 CPU hours). Second, data on 20

small proteins (< 120 residues) were used. I-TASSER folded four of them with a

Calpha-RMSD < 2.5A. The average Calpha-RMSD of the I-TASSER models was 3.9A,

whereas it was 5.9A using TOUCHSTONE-II software. Finally, 20 non-homologous

small proteins (< 120 residues) were taken from the PDB library. An average

Calpha-RMSD of 3.9A was obtained for the third benchmark, with seven cases having

a Calpha-RMSD < 2.5A.

CONCLUSION: Our simulation results show that I-TASSER can consistently predict

the correct folds and sometimes high-resolution models for small single-domain

proteins. Compared with other ab initio modeling methods such as ROSETTA and

TOUCHSTONE II, the average performance of I-TASSER is either much better or is

similar within a lower computational time. These data, together with the

significant performance of automated I-TASSER server (the Zhang-Server) in the

'free modeling' section of the recent Critical Assessment of Structure Prediction

(CASP)7 experiment, demonstrate new progresses in automated ab initio model

generation. The I-TASSER server is freely available for academic users

http://zhang.bioinformatics.ku.edu/I-TASSER.

DOI: 10.1186/1741-7007-5-17

PMCID: PMC1878469

PMID: 17488521 [Indexed for MEDLINE]

2741. Bioinformatics. 2007 May 1;23(9):1159-60. Epub 2007 Mar 1.

NucPred--predicting nuclear localization of proteins.

Brameier M(1), Krings A, MacCallum RM.

Author information:

(1)Bioinformatics Research Center (BiRC), University of Aarhus, Aarhus C,

Denmark.

NucPred analyzes patterns in eukaryotic protein sequences and predicts if a

protein spends at least some time in the nucleus or no time at all. Subcellular

location of proteins represents functional information, which is important for

understanding protein interactions, for the diagnosis of human diseases and for

drug discovery. NucPred is a novel web tool based on regular expression matching

and multiple program classifiers induced by genetic programming. A likelihood

score is derived from the programs for each input sequence and each residue

position. Different forms of visualization are provided to assist the detection

of nuclear localization signals (NLSs). The NucPred server also provides access

to additional sources of biological information (real and predicted) for a better

validation and interpretation of results.AVAILABILITY: The web interface to the

NucPred tool is provided at http://www.sbc.su.se/~maccallr/nucpred. In addition,

the Perl code is made freely available under the GNU Public Licence (GPL) for

simple incorporation into other tools and web servers.

DOI: 10.1093/bioinformatics/btm066

PMID: 17332022 [Indexed for MEDLINE]

2742. J Comput Chem. 2007 May;28(7):1290-305.

Vibalizer: a free, web-based tool for rapid, quantitative comparison and analysis

of calculated vibrational modes.

Grafton AK(1).

Author information:

(1)Division of Science, Lyon College, P.O. Box 2317, Batesville, Arkansas 72503,

USA. akgrafton@lyon.edu

This report describes the development and applications of a software package

called Vibalizer, the first and only method that provides free, fast,

interactive, and quantitative comparison and analysis of calculated vibrational

modes. Using simple forms and menus in a web-based interface, Vibalizer permits

the comparison of vibrational modes from different, but similar molecules and

also performs rapid calculation and comparison of isotopically substituted

molecules' normal modes. Comparing and matching complex vibrational modes can be

completed in seconds with Vibalizer, whereas matching vibrational modes manually

can take hours and gives only qualitative comparisons subject to human error and

differing individual judgments. In addition to these core features, Vibalizer

also provides several other useful features, including the ability to

automatically determine first-approximation mode descriptions, to help users

analyze the results of vibrational frequency calculations. Because the software

can be dimensioned to handle almost arbitrarily large systems, Vibalizer may be

of particular use when analyzing the vibrational modes of complex systems such as

proteins and extended materials systems. Additionally, the ease of use of the

Vibalizer interface and the straightforward interpretation of results may find

favor with educators who incorporate molecular modeling into their classrooms.

The Vibalizer interface is available for free use at http://www.compchem.org, and

it is also available as a locally-installable package that will run on a

Linux-based web server.

DOI: 10.1002/jcc.20642

PMID: 17299728

2743. J Proteome Res. 2007 May;6(5):1728-34. Epub 2007 Mar 31.

Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular

location prediction by incorporating multiple sites.

Chou KC(1), Shen HB.

Author information:

(1)Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego,

California 92130, USA. kchou@ san.rr.com

One of the critical challenges in predicting protein subcellular localization is

how to deal with the case of multiple location sites. Unfortunately, so far, no

efforts have been made in this regard except for the one focused on the proteins

in budding yeast only. For most existing predictors, the multiple-site proteins

are either excluded from consideration or assumed even not existing. Actually,

proteins may simultaneously exist at, or move between, two or more different

subcellular locations. For instance, according to the Swiss-Prot database

(version 50.7, released 19-Sept-2006), among the 33,925 eukaryotic protein

entries that have experimentally observed subcellular location annotations, 2715

have multiple location sites, meaning about 8% bearing the multiplex feature.

Proteins with multiple locations or dynamic feature of this kind are particularly

interesting because they may have some very special biological functions

intriguing to investigators in both basic research and drug discovery. Meanwhile,

according to the same Swiss-Prot database, the number of total eukaryotic protein

entries (except those annotated with "fragment" or those with less than 50 amino

acids) is 90,909, meaning a gap of (90,909-33,925) = 56,984 entries for which no

knowledge is available about their subcellular locations. Although one can use

the computational approach to predict the desired information for the blank, so

far, all the existing methods for predicting eukaryotic protein subcellular

localization are limited in the case of single location site only. To overcome

such a barrier, a new ensemble classifier, named Euk-mPLoc, was developed that

can be used to deal with the case of multiple location sites as well. Euk-mPLoc

is freely accessible to the public as a Web server at

http://202.120.37.186/bioinf/euk-multi. Meanwhile, to support the people working

in the relevant areas, Euk-mPLoc has been used to identify all eukaryotic protein

entries in the Swiss-Prot database that do not have subcellular location

annotations or are annotated as being uncertain. The large-scale results thus

obtained have been deposited at the same Web site via a downloadable file

prepared with Microsoft Excel and named "Tab\_Euk-mPLoc.xls". Furthermore, to

include new entries of eukaryotic proteins and reflect the continuous development

of Euk-mPLoc in both the coverage scope and prediction accuracy, we will timely

update the downloadable file as well as the predictor, and keep users informed by

publishing a short note in the Journal and making an announcement in the Web

Page.

DOI: 10.1021/pr060635i

PMID: 17397210 [Indexed for MEDLINE]

2744. Protein Sci. 2007 May;16(5):947-55.

DDOMAIN: Dividing structures into domains using a normalized domain-domain

interaction profile.

Zhou H(1), Xue B, Zhou Y.

Author information:

(1)Howard Hughes Medical Institute Center for Single Molecule Biophysics,

Department of Physiology and Biophysics, State University of New York at Buffalo,

Buffalo, New York 14214, USA.

Dividing protein structures into domains is proven useful for more accurate

structural and functional characterization of proteins. Here, we develop a

method, called DDOMAIN, that divides structure into DOMAINs using a normalized

contact-based domain-domain interaction profile. Results of DDOMAIN are compared

to AUTHORS annotations (domain definitions are given by the authors who solved

protein structures), as well as to popular SCOP and CATH annotations by human

experts and automatic programs. DDOMAIN's automatic annotations are most

consistent with the AUTHORS annotations (90% agreement in number of domains and

88% agreement in both number of domains and at least 85% overlap in domain

assignment of residues) if its three adjustable parameters are trained by the

AUTHORS annotations. By comparison, the agreement is 83% (81% with at least 85%

overlap criterion) between SCOP-trained DDOMAIN and SCOP annotations and 77%

(73%) between CATH-trained DDOMAIN and CATH annotations. The agreement between

DDOMAIN and AUTHORS annotations goes beyond single-domain proteins (97%, 82%, and

56% for single-, two-, and three-domain proteins, respectively). For an "easy"

data set of proteins whose CATH and SCOP annotations agree with each other in

number of domains, the agreement is 90% (89%) between "easy-set"-trained DDOMAIN

and CATH/SCOP annotations. The consistency between SCOP-trained DDOMAIN and SCOP

annotations is superior to two other recently developed, SCOP-trained, automatic

methods PDP (protein domain parser), and DomainParser 2. We also tested a simple

consensus method made of PDP, DomainParser 2, and DDOMAIN and a different version

of DDOMAIN based on a more sophisticated statistical energy function. The DDOMAIN

server and its executable are available in the services section on

http://sparks.informatics.iupui.edu.

DOI: 10.1110/ps.062597307

PMCID: PMC2206635

PMID: 17456745 [Indexed for MEDLINE]

2745. Biochem Biophys Res Commun. 2007 Apr 20;355(4):1006-11. Epub 2007 Feb 23.

Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular

location prediction by incorporating samples with multiple sites.

Shen HB(1), Chou KC.

Author information:

(1)Department of Biological Chemistry & Molecular Pharmacology, Harvard Medical

School, Boston, MA 02115, USA. hbshen@crystal.harvard.edu

Proteins may simultaneously exist at, or move between, two or more different

subcellular locations. Proteins with multiple locations or dynamic feature of

this kind are particularly interesting because they may have some very special

biological functions intriguing to investigators in both basic research and drug

discovery. For instance, among the 6408 human protein entries that have

experimentally observed subcellular location annotations in the Swiss-Prot

database (version 50.7, released 19-Sept-2006), 973 ( approximately 15%) have

multiple location sites. The number of total human protein entries (except those

annotated with "fragment" or those with less than 50 amino acids) in the same

database is 14,370, meaning a gap of (14,370-6408)=7962 entries for which no

knowledge is available about their subcellular locations. Although one can use

the computational approach to predict the desired information for the gap, so far

all the existing methods for predicting human protein subcellular localization

are limited in the case of single location site only. To overcome such a barrier,

a new ensemble classifier, named Hum-mPLoc, was developed that can be used to

deal with the case of multiple location sites as well. Hum-mPLoc is freely

accessible to the public as a web server at

http://202.120.37.186/bioinf/hum-multi. Meanwhile, for the convenience of people

working in the relevant areas, Hum-mPLoc has been used to identify all human

protein entries in the Swiss-Prot database that do not have subcellular location

annotations or are annotated as being uncertain. The large-scale results thus

obtained have been deposited in a downloadable file prepared with Microsoft Excel

and named "Tab\_Hum-mPLoc.xls". This file is available at the same website and

will be updated twice a year to include new entries of human proteins and reflect

the continuous development of Hum-mPLoc.

DOI: 10.1016/j.bbrc.2007.02.071

PMID: 17346678 [Indexed for MEDLINE]

2746. BMC Infect Dis. 2007 Apr 20;7:32.

epiPATH: an information system for the storage and management of molecular

epidemiology data from infectious pathogens.

Amadoz A(1), González-Candelas F.

Author information:

(1)Institut Cavanilles de Biodiversitat i Biologia Evolutiva and Departament de

Genètica, Universitat de València, Spain. alicia.amadoz@uv.es

<alicia.amadoz@uv.es>

BACKGROUND: Most research scientists working in the fields of molecular

epidemiology, population and evolutionary genetics are confronted with the

management of large volumes of data. Moreover, the data used in studies of

infectious diseases are complex and usually derive from different institutions

such as hospitals or laboratories. Since no public database scheme incorporating

clinical and epidemiological information about patients and molecular information

about pathogens is currently available, we have developed an information system,

composed by a main database and a web-based interface, which integrates both

types of data and satisfies requirements of good organization, simple

accessibility, data security and multi-user support.

RESULTS: From the moment a patient arrives to a hospital or health centre until

the processing and analysis of molecular sequences obtained from infectious

pathogens in the laboratory, lots of information is collected from different

sources. We have divided the most relevant data into 12 conceptual modules around

which we have organized the database schema. Our schema is very complete and it

covers many aspects of sample sources, samples, laboratory processes, molecular

sequences, phylogenetics results, clinical tests and results, clinical

information, treatments, pathogens, transmissions, outbreaks and bibliographic

information. Communication between end-users and the selected Relational Database

Management System (RDMS) is carried out by default through a command-line window

or through a user-friendly, web-based interface which provides access and

management tools for the data.

CONCLUSION: epiPATH is an information system for managing clinical and molecular

information from infectious diseases. It facilitates daily work related to

infectious pathogens and sequences obtained from them. This software is intended

for local installation in order to safeguard private data and provides advanced

SQL-users the flexibility to adapt it to their needs. The database schema, tool

scripts and web-based interface are free software but data stored in our database

server are not publicly available. epiPATH is distributed under the terms of GNU

General Public License. More details about epiPATH can be found at

http://genevo.uv.es/epipath.

DOI: 10.1186/1471-2334-7-32

PMCID: PMC1868736

PMID: 17448245 [Indexed for MEDLINE]

2747. BMC Bioinformatics. 2007 Apr 19;8:129.

CGKB: an annotation knowledge base for cowpea (Vigna unguiculata L.) methylation

filtered genomic genespace sequences.

Chen X(1), Laudeman TW, Rushton PJ, Spraggins TA, Timko MP.

Author information:

(1)Department of Microbiology, University of Virginia Health System,

Charlottesville, VA 29908, USA. xc3m@virginia.edu <xc3m@virginia.edu>

BACKGROUND: Cowpea [Vigna unguiculata (L.) Walp.] is one of the most important

food and forage legumes in the semi-arid tropics because of its ability to

tolerate drought and grow on poor soils. It is cultivated mostly by poor farmers

in developing countries, with 80% of production taking place in the dry savannah

of tropical West and Central Africa. Cowpea is largely an underexploited crop

with relatively little genomic information available for use in applied plant

breeding. The goal of the Cowpea Genomics Initiative (CGI), funded by the

Kirkhouse Trust, a UK-based charitable organization, is to leverage modern

molecular genetic tools for gene discovery and cowpea improvement. One aspect of

the initiative is the sequencing of the gene-rich region of the cowpea genome

(termed the genespace) recovered using methylation filtration technology and

providing annotation and analysis of the sequence data.

DESCRIPTION: CGKB, Cowpea Genespace/Genomics Knowledge Base, is an annotation

knowledge base developed under the CGI. The database is based on information

derived from 298,848 cowpea genespace sequences (GSS) isolated by methylation

filtering of genomic DNA. The CGKB consists of three knowledge bases: GSS

annotation and comparative genomics knowledge base, GSS enzyme and metabolic

pathway knowledge base, and GSS simple sequence repeats (SSRs) knowledge base for

molecular marker discovery. A homology-based approach was applied for annotations

of the GSS, mainly using BLASTX against four public FASTA formatted protein

databases (NCBI GenBank Proteins, UniProtKB-Swiss-Prot, UniprotKB-PIR (Protein

Information Resource), and UniProtKB-TrEMBL). Comparative genome analysis was

done by BLASTX searches of the cowpea GSS against four plant proteomes from

Arabidopsis thaliana, Oryza sativa, Medicago truncatula, and Populus trichocarpa.

The possible exons and introns on each cowpea GSS were predicted using the

HMM-based Genscan gene predication program and the potential domains on annotated

GSS were analyzed using the HMMER package against the Pfam database. The

annotated GSS were also assigned with Gene Ontology annotation terms and

integrated with 228 curated plant metabolic pathways from the Arabidopsis

Information Resource (TAIR) knowledge base. The UniProtKB-Swiss-Prot ENZYME

database was used to assign putative enzymatic function to each GSS. Each GSS was

also analyzed with the Tandem Repeat Finder (TRF) program in order to identify

potential SSRs for molecular marker discovery. The raw sequence data, processed

annotation, and SSR results were stored in relational tables designed in

key-value pair fashion using a PostgreSQL relational database management system.

The biological knowledge derived from the sequence data and processed results are

represented as views or materialized views in the relational database management

system. All materialized views are indexed for quick data access and retrieval.

Data processing and analysis pipelines were implemented using the Perl

programming language. The web interface was implemented in JavaScript and Perl

CGI running on an Apache web server. The CPU intensive data processing and

analysis pipelines were run on a computer cluster of more than 30 dual-processor

Apple XServes. A job management system called Vela was created as a robust way to

submit large numbers of jobs to the Portable Batch System (PBS).

CONCLUSION: CGKB is an integrated and annotated resource for cowpea GSS with

features of homology-based and HMM-based annotations, enzyme and pathway

annotations, GO term annotation, toolkits, and a large number of other facilities

to perform complex queries. The cowpea GSS, chloroplast sequences, mitochondrial

sequences, retroelements, and SSR sequences are available as FASTA formatted

files and downloadable at CGKB. This database and web interface are publicly

accessible at http://cowpeagenomics.med.virginia.edu/CGKB/.

DOI: 10.1186/1471-2105-8-129

PMCID: PMC1868039

PMID: 17445272 [Indexed for MEDLINE]

2748. Bioinformatics. 2007 Apr 15;23(8):942-9. Epub 2007 Mar 24.

POPI: predicting immunogenicity of MHC class I binding peptides by mining

informative physicochemical properties.

Tung CW(1), Ho SY.

Author information:

(1)Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan.

MOTIVATION: Both modeling of antigen-processing pathway including major

histocompatibility complex (MHC) binding and immunogenicity prediction of those

MHC-binding peptides are essential to develop a computer-aided system of

peptide-based vaccine design that is one goal of immunoinformatics. Numerous

studies have dealt with modeling the immunogenic pathway but not the intractable

problem of immunogenicity prediction due to complex effects of many intrinsic and

extrinsic factors. Moderate affinity of the MHC-peptide complex is essential to

induce immune responses, but the relationship between the affinity and peptide

immunogenicity is too weak to use for predicting immunogenicity. This study

focuses on mining informative physicochemical properties from known experimental

immunogenicity data to understand immune responses and predict immunogenicity of

MHC-binding peptides accurately.

RESULTS: This study proposes a computational method to mine a feature set of

informative physicochemical properties from MHC class I binding peptides to

design a support vector machine (SVM) based system (named POPI) for the

prediction of peptide immunogenicity. High performance of POPI arises mainly from

an inheritable bi-objective genetic algorithm, which aims to automatically

determine the best number m out of 531 physicochemical properties, identify these

m properties and tune SVM parameters simultaneously. The dataset consisting of

428 human MHC class I binding peptides belonging to four classes of

immunogenicity was established from MHCPEP, a database of MHC-binding peptides

(Brusic et al., 1998). POPI, utilizing the m = 23 selected properties, performs

well with the accuracy of 64.72% using leave-one-out cross-validation, compared

with two sequence alignment-based prediction methods ALIGN (54.91%) and PSI-BLAST

(53.23%). POPI is the first computational system for prediction of peptide

immunogenicity based on physicochemical properties.

AVAILABILITY: A web server for prediction of peptide immunogenicity (POPI) and

the used dataset of MHC class I binding peptides (PEPMHCI) are available at

http://iclab.life.nctu.edu.tw/POPI

DOI: 10.1093/bioinformatics/btm061

PMID: 17384427 [Indexed for MEDLINE]

2749. Bioinformatics. 2007 Apr 15;23(8):1026-8. Epub 2007 Feb 19.

Gepard: a rapid and sensitive tool for creating dotplots on genome scale.

Krumsiek J(1), Arnold R, Rattei T.

Author information:

(1)Institute for Bioinformatics (MIPS), GSF National Research Center for

Environment and Health, Neuherberg, Germany. krumsiek@in.tum.de

Gepard provides a user-friendly, interactive application for the quick creation

of dotplots. It utilizes suffix arrays to reduce the time complexity of dotplot

calculation to Theta(m\*log n). A client-server mode, which is a novel feature for

dotplot creation software, allows the user to calculate dotplots and color them

by functional annotation without any prior downloading of sequence or annotation

data.AVAILABILITY: Both source codes and executable binaries are available at

http://mips.gsf.de/services/analysis/gepard

DOI: 10.1093/bioinformatics/btm039

PMID: 17309896 [Indexed for MEDLINE]

2750. Bioinformatics. 2007 Apr 15;23(8):1032-4. Epub 2007 Feb 18.

ClusterDraw web server: a tool to identify and visualize clusters of binding

motifs for transcription factors.

Papatsenko D(1).

Author information:

(1)Department of Molecular and Cell Biology, University of California, Berkeley,

Berkeley, CA 94720, USA. dxp@berkeley.edu

ClusterDraw is a program aimed to identification of binding sites and

binding-site clusters. Major difference of the ClusterDraw from existing tools is

its ability to scan a wide range of parameter values and weigh statistical

significance of all possible clusters, smaller than a selected size. The program

produces graphs along with decorated FASTA files. ClusterDraw web server is

available at the following URL: http://flydev.berkeley.edu/cgi-bin/cld/submit.cgi

DOI: 10.1093/bioinformatics/btm047

PMID: 17308342 [Indexed for MEDLINE]

2751. BMC Bioinformatics. 2007 Apr 3;8:114.

From genes to functional classes in the study of biological systems.

Al-Shahrour F(1), Arbiza L, Dopazo H, Huerta-Cepas J, Mínguez P, Montaner D,

Dopazo J.

Author information:

(1)Bioinformatics Department, Valencia, Spain. falshahrour@cipf.es

<falshahrour@cipf.es>

BACKGROUND: With the popularization of high-throughput techniques, the need for

procedures that help in the biological interpretation of results has increased

enormously. Recently, new procedures inspired in systems biology criteria have

started to be developed.

RESULTS: Here we present FatiScan, a web-based program which implements a

threshold-independent test for the functional interpretation of large-scale

experiments that does not depend on the pre-selection of genes based on the

multiple application of independent tests to each gene. The test implemented aims

to directly test the behaviour of blocks of functionally related genes, instead

of focusing on single genes. In addition, the test does not depend on the type of

the data used for obtaining significance values, and consequently different types

of biologically informative terms (gene ontology, pathways, functional motifs,

transcription factor binding sites or regulatory sites from CisRed) can be

applied to different classes of genome-scale studies. We exemplify its

application in microarray gene expression, evolution and interactomics.

CONCLUSION: Methods for gene set enrichment which, in addition, are independent

from the original data and experimental design constitute a promising alternative

for the functional profiling of genome-scale experiments. A web server that

performs the test described and other similar ones can be found at:

http://www.babelomics.org.

DOI: 10.1186/1471-2105-8-114

PMCID: PMC1853114

PMID: 17407596 [Indexed for MEDLINE]

2752. Bioinformatics. 2007 Apr 1;23(7):895-7. Epub 2007 Feb 5.

NetPhosYeast: prediction of protein phosphorylation sites in yeast.

Ingrell CR(1), Miller ML, Jensen ON, Blom N.

Author information:

(1)University of Southern Denmark, Campusvej 55, DK-5230, Odense M, Denmark.

We here present a neural network-based method for the prediction of protein

phosphorylation sites in yeast--an important model organism for basic research.

Existing protein phosphorylation site predictors are primarily based on mammalian

data and show reduced sensitivity on yeast phosphorylation sites compared to

those in humans, suggesting the need for an yeast-specific phosphorylation site

predictor. NetPhosYeast achieves a correlation coefficient close to 0.75 with a

sensitivity of 0.84 and specificity of 0.90 and outperforms existing predictors

in the identification of phosphorylation sites in yeast.AVAILABILITY: The

NetPhosYeast prediction service is available as a public web server at

http://www.cbs.dtu.dk/services/NetPhosYeast/.

DOI: 10.1093/bioinformatics/btm020

PMID: 17282998 [Indexed for MEDLINE]

2753. Bioinformatics. 2007 Apr 1;23(7):901-2. Epub 2007 Feb 3.

DFprot: a webtool for predicting local chain deformability.

Garzón JI(1), Kovacs J, Abagyan R, Chacón P.

Author information:

(1)Centro de Investigaciones Biológicas, CSIC, Ramiro de Maeztu 9, 28040 Madrid,

Spain.

DFprot is a web-based server for predicting main-chain deformability from a

single protein conformation. The server automatically performs a normal-mode

analysis (NMA) of the uploaded structure and calculates its capability to deform

at each of its residues. Non-specialists can easily and rapidly obtain a

quantitative first approximation of the flexibility of their structures with a

simple and efficient interface.AVAILABILITY:

http://sbg.cib.csic.es/Software/DFprot.

DOI: 10.1093/bioinformatics/btm014

PMID: 17277334 [Indexed for MEDLINE]

2754. Bioinformatics. 2007 Apr 1;23(7):802-8. Epub 2007 Jan 31.

PROMALS: towards accurate multiple sequence alignments of distantly related

proteins.

Pei J(1), Grishin NV.

Author information:

(1)Howard Hughes Medical Institute, The University of Texas Southwestern Medical

Center at Dallas, 6001 Forest Park Road, Dallas, TX 75390-9050, USA.

jpei@chop.swmed.edu

MOTIVATION: Accurate multiple sequence alignments are essential in protein

structure modeling, functional prediction and efficient planning of experiments.

Although the alignment problem has attracted considerable attention, preparation

of high-quality alignments for distantly related sequences remains a difficult

task.

RESULTS: We developed PROMALS, a multiple alignment method that shows promising

results for protein homologs with sequence identity below 10%, aligning close to

half of the amino acid residues correctly on average. This is about three times

more accurate than traditional pairwise sequence alignment methods. PROMALS

algorithm derives its strength from several sources: (i) sequence database

searches to retrieve additional homologs; (ii) accurate secondary structure

prediction; (iii) a hidden Markov model that uses a novel combined scoring of

amino acids and secondary structures; (iv) probabilistic consistency-based

scoring applied to progressive alignment of profiles. Compared to the best

alignment methods that do not use secondary structure prediction and database

searches (e.g. MUMMALS, ProbCons and MAFFT), PROMALS is up to 30% more accurate,

with improvement being most prominent for highly divergent homologs. Compared to

SPEM and HHalign, which also employ database searches and secondary structure

prediction, PROMALS shows an accuracy improvement of several percent.

AVAILABILITY: The PROMALS web server is available at:

http://prodata.swmed.edu/promals/.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btm017

PMID: 17267437 [Indexed for MEDLINE]

2755. Curr Protein Pept Sci. 2007 Apr;8(2):181-8.

Computer-assisted protein domain boundary prediction using the DomPred server.

Bryson K(1), Cozzetto D, Jones DT.

Author information:

(1)Department of Computer Science, University College London, Gower Street,

London WC1E 6BT, United Kingdom.

Domain prediction from sequence is a particularly challenging task, and

currently, a large variety of different methodologies are employed to tackle the

task. Here we try to classify these diverse approaches into a number of broad

categories. Completely automatic domain prediction from sequence alone is

currently fraught with problems, but this should not be so surprising since human

experts currently have significant disagreement on domain assignment even when

given the structures. It can be argued that we should only test the domain

prediction methods on benchmark data that human experts agree upon and this is

the approach we take in this paper. Even for the data sets on which human experts

agree, automatic structure-based domain assignment still cannot always agree, and

so again it is still unlikely that domain prediction methods will reliably obtain

correct results completely automatically. We make the argument that

computer-assisted domain prediction is a more achievable goal. With this aim in

mind, we present the DomPred server. This server provides the user with the

results from two completely different categories of method (DPS and DomSSEA). In

this paper, each method is individually benchmarked against one of the latest

domain prediction benchmarks to provide information about their respective

reliabilities. A variety of different benchmark scores are employed since the

accuracy of a domain prediction method depends critically on what types of

results one wishes to obtain (single/multi-domain classification, domain number,

residue linker positions, etc.). Also both of these methods, implemented within

the DomPred server, can suggest alternative domain predictions, allowing the user

to make the final decision based on these results and applying their own

background knowledge to the problem. The DomPred server is available from the

URL:http://bioinf.cs.ucl.ac.uk/software.html.

PMID: 17430199 [Indexed for MEDLINE]

2756. Fungal Genet Biol. 2007 Apr;44(4):231-41. Epub 2007 Jan 10.

Separation of sequences from host-pathogen interface using triplet nucleotide

frequencies.

Emmersen J(1), Rudd S, Mewes HW, Tetko IV.

Author information:

(1)Institut for Miljø og Bioteknologi, Aalborg Universitet, Sohngaardsholmsvej

49, 9000 Aalborg, Denmark.

The identification of genes involved in host-pathogen interactions is important

for the elucidation of mechanisms of disease resistance and host susceptibility.

A traditional way to classify the origin of genes sampled from a pool of mixed

cDNA is through sequence similarity to known genes from either the pathogen or

host organism or other closely related species. This approach does not work when

the identified sequence has no close homologues in the sequence databases. In our

previous studies, we classified genes using their codon frequencies. This method,

however, explicitly required the prediction of CDS regions and thus could not be

applied to sequences composed from the non-coding regions of genes. In this

study, we show that the use of sliding-window triplet frequencies extends the

application of the algorithm to both coding and non-coding sequences and also

increases the prediction accuracy of a Support Vector Machine classifier from

95.6+/-0.3 to 96.5+/-0.2. Thus the use of the triplet frequencies increased the

prediction accuracy of the new method by more than 20% compared to our previous

approach. A functional analysis of sequences detected gene families having

significantly higher or lower probability to be correctly classified compared to

the average accuracy of the method is described. The server to perform

classification of EST sequences using triplet frequencies is available at (URL:

http://mips.gsf.de/proj/est3).

DOI: 10.1016/j.fgb.2006.11.010

PMID: 17218127 [Indexed for MEDLINE]

2757. Hum Pathol. 2007 Apr;38(4):546-54. Epub 2007 Jan 31.

Feasibility and diagnostic agreement in teledermatopathology using a virtual

slide system.

Massone C(1), Soyer HP, Lozzi GP, Di Stefani A, Leinweber B, Gabler G, Asgari M,

Boldrini R, Bugatti L, Canzonieri V, Ferrara G, Kodama K, Mehregan D, Rongioletti

F, Janjua SA, Mashayekhi V, Vassilaki I, Zelger B, Zgavec B, Cerroni L, Kerl H.

Author information:

(1)Department of Dermatology, Medical University of Graz, A-8036 Graz, Austria.

We investigated the feasibility and diagnostic agreement of a virtual slide

system (VSS) in teledermatopathology. Forty-six biopsy specimens from

inflammatory skin diseases were selected and scanned with a VSS at the Research

Unit of Teledermatology, Medical University of Graz, Graz, Austria. Images were

stored on a virtual slide server on which a specific Web application suited for

telepathology (http://telederm.org/research/dermatopath/) runs. Twelve

teleconsultants from 6 different countries reviewed the 46 cases, working

directly on the Web application. Telediagnoses agreed with gold standard and

conventional diagnosis with an average of 73% and 74%, respectively. Complete

concordance among all teleconsultants with gold standard and conventional

diagnosis was found in 20% of the cases. In 10 cases in which complete clinical

data were missing, the average agreement of telediagnosis with gold standard

diagnosis and conventional diagnosis decreased to 65% and 66%, respectively. Only

3 of 4 cases of inflammatory skin diseases were correctly diagnosed remotely with

VSS. The system that we have used, despite its usability, is not completely

feasible for teledermatopathology of inflammatory skin disease. Moreover, the

performance seems to have been influenced by the availability of complete

clinical data and by the intrinsic difficulty of the pathology of inflammatory

skin diseases.

DOI: 10.1016/j.humpath.2006.10.006

PMID: 17270240 [Indexed for MEDLINE]

2758. Neuroinformatics. 2007 Spring;5(1):35-58.

Toward a workbench for rodent brain image data: systems architecture and design.

Moene IA(1), Subramaniam S, Darin D, Leergaard TB, Bjaalie JG.

Author information:

(1)Neural Systems and Graphics Computing Laboratory, Centre for Molecular Biology

and Neuroscience & Institute of Basic Medical Sciences, University of Oslo, PO

Box 1105 Blindern, N-0317 Oslo, Norway.

We present a novel system for storing and manipulating microscopic images from

sections through the brain and higher-level data extracted from such images. The

system is designed and built on a three-tier paradigm and provides the research

community with a web-based interface for facile use in neuroscience research. The

Oracle relational database management system provides the ability to store a

variety of objects relevant to the images and provides the framework for complex

querying of data stored in the system. Further, the suite of applications

intimately tied into the infrastructure in the application layer provide the user

the ability not only to query and visualize the data, but also to perform

analysis operations based on the tools embedded into the system. The presentation

layer uses extant protocols of the modern web browser and this provides ease of

use of the system. The present release, named Functional Anatomy of the

Cerebro-Cerebellar System (FACCS), available through The Rodent Brain Workbench

(http:// rbwb.org/), is targeted at the functional anatomy of the

cerebro-cerebellar system in rats, and holds axonal tracing data from these

projections. The system is extensible to other circuits and projections and to

other categories of image data and provides a unique environment for analysis of

rodent brain maps in the context of anatomical data. The FACCS application

assumes standard animal brain atlas models and can be extended to future models.

The system is available both for interactive use from a remote web-browser client

as well as for download to a local server machine.

PMID: 17426352 [Indexed for MEDLINE]

2759. Proteomics. 2007 Apr;7(7):1117-20.

Yeast proteome map (update 2006).

Perrot M(1), Guieysse-Peugeot AL, Massoni A, Espagne C, Claverol S, Silva RM,

Jenö P, Santos M, Bonneu M, Boucherie H.

Author information:

(1)Institut de Biochemie et Génétique Cellulaires, UMR CNRS 5095, Bordeaux,

France.

To improve the potential of two-dimensional gel electrophoresis for proteomic

investigations in yeast we have undertaken the systematic identification of

Saccharomyces cerevisiae proteins separated on 2-D gels. We report here the

identification of 187 novel protein spots. They were identified by two methods,

mass spectrometry and gene inactivation. These identifications extend the number

of protein spots identified on our yeast 2-D proteome map to 602, i.e. nearly

half the detectable spots of the proteome map. These spots correspond to 417

different proteins. The reference map and the list of identified proteins can be

accessed on the Yeast Protein Map server (www.ibgc.u-bordeaux2.fr/YPM).

DOI: 10.1002/pmic.200600952

PMID: 17351888 [Indexed for MEDLINE]

2760. Biochem Biophys Res Commun. 2007 Mar 9;354(2):498-504. Epub 2007 Jan 12.

Construction of mathematical model for high-level expression of foreign genes in

pPIC9 vector and its verification.

Wu B(1), Cha L, Du Z, Ying X, Li H, Xu L, Zheng X, Li E, Li W.

Author information:

(1)Center of Computational Biology, Beijing Institute of Basic Medical Sciences,

Beijing 100850, China.

In this report, we introduced a mathematical model for high-level expression of

foreign genes in pPIC9 vector. At first, we collected 40 heterologous genes

expressed in pPIC9 vector, and these 40 genes were classified into high-level

expression group (expression level >100mg/L, 12 genes) and low-level expression

group (expression level <100mg/L, 28 genes). Then, the Naive Bayes method was

used to construct the model with RNA secondary structure profile of 3'-end of

foreign genes as features. The classification accuracy from leave-one-out

cross-validation was 100%. Finally, another five genes collected from literatures

were used to test the ability of the model. The results indicated that there were

four genes correctly predicted. In addition, the model was also verified by

expressing human neutrophil gelatinase-associated lipocalin (NGAL) gene with

expression level more than 100mg/L. Therefore, we propose that the model can be

used to predict the expression level of heterologous genes before experiments and

optimize the experiment designs to obtain the high-level expression. Furthermore,

we have developed a web server for evaluation and design for high-level

expression of foreign genes, which is accessible at

http://ppic9.med.stu.edu.cn/ppic9.

DOI: 10.1016/j.bbrc.2007.01.002

PMID: 17239823 [Indexed for MEDLINE]

2761. BMC Bioinformatics. 2007 Mar 8;8 Suppl 1:S6.

The MEPS server for identifying protein conformational epitopes.

Castrignanò T(1), De Meo PD, Carrabino D, Orsini M, Floris M, Tramontano A.

Author information:

(1)CASPUR, Consorzio Interuniversitario per le Applicazioni di Supercalcolo per

Universita' e Ricerca, Via dei Tizii, 6/b, 5 I-00185 Rome, Italy.

Tiziana.Castrignano@caspur.it

BACKGROUND: One of the most interesting problems in molecular immunology is

epitope mapping, i.e. the identification of the regions of interaction between an

antigen and an antibody. The solution to this problem, even if approximate, would

help in designing experiments to precisely map the residues involved in the

interaction and could be instrumental both in designing peptides able to mimic

the interacting surface of the antigen and in understanding where immunologically

important regions are located in its three-dimensional structure. From an

experimental point of view, both genetically encoded and chemically synthesised

peptide libraries can be used to identify sequences recognized by a given

antibody. The problem then arises of which region of a folded protein the

selected peptides correspond to.

RESULTS: We have developed a method able to find the surface region of a protein

that can be effectively mimicked by a peptide, given the structure of the protein

and the maximum number of side chains deemed to be required for recognition. The

method is implemented as a publicly available server. It can also find and report

all peptide sequences of a specified length that can mimic the surface of a given

protein and store them in a database. The immediate application of the server is

the mapping of antibody epitopes, however the system is sufficiently flexible for

allowing other questions to be asked, for example one can compare the peptides

representing the surface of two proteins known to interact with the same

macromolecule to find which is the most likely interacting region.

CONCLUSION: We believe that the MEPS server, available at

http://www.caspur.it/meps, will be a useful tool for immunologists and structural

and computational biologists. We plan to use it ourselves to implement a database

of "surface mimicking peptides" for all proteins of known structure and proteins

that can be reliably modelled by comparative modelling.

DOI: 10.1186/1471-2105-8-S1-S6

PMCID: PMC1885858

PMID: 17430573 [Indexed for MEDLINE]

2762. BMC Bioinformatics. 2007 Mar 8;8 Suppl 1:S3.

A computational approach for detecting peptidases and their specific inhibitors

at the genome level.

Bartoli L(1), Calabrese R, Fariselli P, Mita DG, Casadio R.

Author information:

(1)Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna,

Bologna, Italy. lisa@biocomp.unibo.it

BACKGROUND: Peptidases are proteolytic enzymes responsible for fundamental

cellular activities in all organisms. Apparently about 2-5% of the genes encode

for peptidases, irrespectively of the organism source. The basic peptidase

function is "protein digestion" and this can be potentially dangerous in living

organisms when it is not strictly controlled by specific inhibitors. In genome

annotation a basic question is to predict gene function. Here we describe a

computational approach that can filter peptidases and their inhibitors out of a

given proteome. Furthermore and as an added value to MEROPS, a specific database

for peptidases already available in the public domain, our method can predict

whether a pair of peptidase/inhibitor can interact, eventually listing all

possible predicted ligands (peptidases and/or inhibitors).

RESULTS: We show that by adopting a decision-tree approach the accuracy of

PROSITE and HMMER in detecting separately the four major peptidase types (Serine,

Aspartic, Cysteine and Metallo- Peptidase) and their inhibitors among a non

redundant set of globular proteins can be improved by some percentage points with

respect to that obtained with each method separately. More importantly, our

method can then predict pairs of peptidases and interacting inhibitors, scoring a

joint global accuracy of 99% with coverage for the positive cases

(peptidase/inhibitor) close to 100% and a correlation coefficient of 0.91%. In

this task the decision-tree approach outperforms the single methods.

CONCLUSION: The decision-tree can reliably classify protein sequences as

peptidases or inhibitors, belonging to a certain class, and can provide a

comprehensive list of possible interacting pairs of peptidase/inhibitor. This

information can help the design of experiments to detect interacting

peptidase/inhibitor complexes and can speed up the selection of possible

interacting candidates, without searching for them separately and manually

combining the obtained results. A web server specifically developed for

annotating peptidases and their inhibitors (HIPPIE) is available at

http://gpcr.biocomp.unibo.it/cgi/predictors/hippie/pred\_hippie.cgi.

DOI: 10.1186/1471-2105-8-S1-S3

PMCID: PMC1885855

PMID: 17430570 [Indexed for MEDLINE]

2763. Bioinformatics. 2007 Mar 1;23(5):634-6. Epub 2007 Jan 19.

DP-Bind: a web server for sequence-based prediction of DNA-binding residues in

DNA-binding proteins.

Hwang S(1), Gou Z, Kuznetsov IB.

Author information:

(1)Gen\*NY\*Sis Center for Excellence in Cancer Genomics, Department of

Epidemiology and Biostatistics, One Discovery Drive, University at Albany,

Rensselaer, NY 12144, USA.

This article describes DP-Bind, a web server for predicting DNA-binding sites in

a DNA-binding protein from its amino acid sequence. The web server implements

three machine learning methods: support vector machine, kernel logistic

regression and penalized logistic regression. Prediction can be performed using

either the input sequence alone or an automatically generated profile of

evolutionary conservation of the input sequence in the form of PSI-BLAST

position-specific scoring matrix (PSSM). PSSM-based kernel logistic regression

achieves the accuracy of 77.2%, sensitivity of 76.4% and specificity of 76.6%.

The outputs of all three individual methods are combined into a consensus

prediction to help identify positions predicted with high level of

confidence.AVAILABILITY: Freely available at http://lcg.rit.albany.edu/dp-bind.

SUPPLEMENTARY INFORMATION:

http://lcg.rit.albany.edu/dp-bind/dpbind\_supplement.html.

DOI: 10.1093/bioinformatics/btl672

PMID: 17237068 [Indexed for MEDLINE]

2764. Bioinformatics. 2007 Mar 1;23(5):538-44. Epub 2007 Jan 19.

Improving the accuracy of transmembrane protein topology prediction using

evolutionary information.

Jones DT(1).

Author information:

(1)Department of Computer Science, University College London, Gower Street,

London WC1E 6BT, United Kingdom. dtj@cs.ucl.ac.uk

MOTIVATION: Many important biological processes such as cell signaling, transport

of membrane-impermeable molecules, cell-cell communication, cell recognition and

cell adhesion are mediated by membrane proteins. Unfortunately, as these proteins

are not water soluble, it is extremely hard to experimentally determine their

structure. Therefore, improved methods for predicting the structure of these

proteins are vital in biological research. In order to improve transmembrane

topology prediction, we evaluate the combined use of both integrated signal

peptide prediction and evolutionary information in a single algorithm.

RESULTS: A new method (MEMSAT3) for predicting transmembrane protein topology

from sequence profiles is described and benchmarked with full cross-validation on

a standard data set of 184 transmembrane proteins. The method is found to predict

both the correct topology and the locations of transmembrane segments for 80% of

the test set. This compares with accuracies of 62-72% for other popular methods

on the same benchmark. By using a second neural network specifically to

discriminate transmembrane from globular proteins, a very low overall false

positive rate (0.5%) can also be achieved in detecting transmembrane proteins.

AVAILABILITY: An implementation of the described method is available both as a

web server (http://www.psipred.net) and as downloadable source code from

http://bioinf.cs.ucl.ac.uk/memsat. Both the server and source code files are free

to non-commercial users. Benchmark and training data are also available from

http://bioinf.cs.ucl.ac.uk/memsat.

DOI: 10.1093/bioinformatics/btl677

PMID: 17237066 [Indexed for MEDLINE]

2765. Bioinformatics. 2007 Mar 1;23(5):637-8. Epub 2007 Jan 19.

SMotif: a server for structural motifs in proteins.

Pugalenthi G(1), Suganthan PN, Sowdhamini R, Chakrabarti S.

Author information:

(1)School of Electrical and Electronic Engineering, Nanyang Technological

University, Singapore 639798.

SMotif is a server that identifies important structural segments or motifs for a

given protein structure(s) based on conservation of both sequential as well as

important structural features such as solvent inaccessibility, secondary

structural content, hydrogen bonding pattern and residue packing. This server

also provides three-dimensional orientation patterns of the identified motifs in

terms of inter-motif distances and torsion angles. These motifs may form the

common core and therefore, can also be employed to design and rationalize protein

engineering and folding experiments.AVAILABILITY: SMotif server is available via

the URL http://caps.ncbs.res.in/SMotif/index.html.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btl679

PMID: 17237055 [Indexed for MEDLINE]

2766. BMC Bioinformatics. 2007 Feb 27;8:65.

AGGRESCAN: a server for the prediction and evaluation of "hot spots" of

aggregation in polypeptides.

Conchillo-Solé O(1), de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S.

Author information:

(1)Institut de Biotecnologia i de Biomedicina (IBB), Universitat Autònoma de

Barcelona, 08193 Bellaterra, Spain. Oscar.Conhillo@bioinf.uab.cat

<Oscar.Conhillo@bioinf.uab.cat>

BACKGROUND: Protein aggregation correlates with the development of several

debilitating human disorders of growing incidence, such as Alzheimer's and

Parkinson's diseases. On the biotechnological side, protein production is often

hampered by the accumulation of recombinant proteins into aggregates. Thus, the

development of methods to anticipate the aggregation properties of polypeptides

is receiving increasing attention. AGGRESCAN is a web-based software for the

prediction of aggregation-prone segments in protein sequences, the analysis of

the effect of mutations on protein aggregation propensities and the comparison of

the aggregation properties of different proteins or protein sets.

RESULTS: AGGRESCAN is based on an aggregation-propensity scale for natural amino

acids derived from in vivo experiments and on the assumption that short and

specific sequence stretches modulate protein aggregation. The algorithm is shown

to identify a series of protein fragments involved in the aggregation of

disease-related proteins and to predict the effect of genetic mutations on their

deposition propensities. It also provides new insights into the differential

aggregation properties displayed by globular proteins, natively unfolded

polypeptides, amyloidogenic proteins and proteins found in bacterial inclusion

bodies.

CONCLUSION: By identifying aggregation-prone segments in proteins, AGGRESCAN

http://bioinf.uab.es/aggrescan/ shall facilitate (i) the identification of

possible therapeutic targets for anti-depositional strategies in conformational

diseases and (ii) the anticipation of aggregation phenomena during storage or

recombinant production of bioactive polypeptides or polypeptide sets.

DOI: 10.1186/1471-2105-8-65

PMCID: PMC1828741

PMID: 17324296 [Indexed for MEDLINE]

2767. Bioinformatics. 2007 Feb 15;23(4):504-6. Epub 2006 Dec 6.

AllerTool: a web server for predicting allergenicity and allergic

cross-reactivity in proteins.

Zhang ZH(1), Koh JL, Zhang GL, Choo KH, Tammi MT, Tong JC.

Author information:

(1)Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613.

zhzhang@i2r.a-star.edu.sg

Assessment of potential allergenicity and patterns of cross-reactivity is

necessary whenever novel proteins are introduced into human food chain. Current

bioinformatic methods in allergology focus mainly on the prediction of allergenic

proteins, with no information on cross-reactivity patterns among known allergens.

In this study, we present AllerTool, a web server with essential tools for the

assessment of predicted as well as published cross-reactivity patterns of

allergens. The analysis tools include graphical representation of allergen

cross-reactivity information; a local sequence comparison tool that displays

information of known cross-reactive allergens; a sequence similarity search tool

for assessment of cross-reactivity in accordance to FAO/WHO Codex alimentarius

guidelines; and a method based on support vector machine (SVM). A 10-fold

cross-validation results showed that the area under the receiver operating curve

(A(ROC)) of SVM models is 0.90 with 86.00% sensitivity (SE) at specificity (SP)

of 86.00%.AVAILABILITY: AllerTool is freely available at

http://research.i2r.a-star.edu.sg/AllerTool/.

DOI: 10.1093/bioinformatics/btl621

PMID: 17150996 [Indexed for MEDLINE]

2768. Bioinformatics. 2007 Feb 15;23(4):498-9. Epub 2006 Dec 4.

AutoGRAPH: an interactive web server for automating and visualizing comparative

genome maps.

Derrien T(1), André C, Galibert F, Hitte C.

Author information:

(1)CNRS UMR6061 Génétique et Développement, Université de Rennes1, IFR140, 2 Av

du Pr. Léon Bernard, CS 34317, 35043, France.

AutoGRAPH is an interactive web server for automatic multi-species comparative

genomics analyses based on personal datasets or pre-inserted public datasets.

This program automatically identifies conserved segments (CS) and breakpoint

regions, assesses the conservation of marker/gene order between organisms,

constructs synteny maps for two to three species and generates high-quality,

interactive displays facilitating the identification of chromosomal

rearrangements. AutoGRAPH can also be used for the integration and comparison of

several types of genomic resources (meiotic maps, radiation hybrid maps and

genome sequences) for a single species, making AutoGRAPH a versatile tool for

comparative genomics analysis.AVAILABILITY:

http://genoweb.univ-rennes1.fr/tom\_dog/AutoGRAPH/.

SUPPLEMENTARY INFORMATION: A description of the algorithm and additional

information are available at

http://genoweb.univ-rennes1.fr/tom\_dog/AutoGRAPH/Tutorial.php.

DOI: 10.1093/bioinformatics/btl618

PMID: 17145741 [Indexed for MEDLINE]

2769. Bioinformatics. 2007 Feb 15;23(4):524-6. Epub 2006 Oct 31.

OmicBrowse: a browser of multidimensional omics annotations.

Toyoda T(1), Mochizuki Y, Player K, Heida N, Kobayashi N, Sakaki Y.

Author information:

(1)Genomic Sciences Center, Yokohama Institute, RIKEN, 1-77-22, Suehiro, Tsurumi,

Yokohama, Kanagawa, Japan. toyop@gsc.riken.go.jp

OmicBrowse is a browser to explore multiple datasets coordinated in the

multidimensional omic space integrating omics knowledge ranging from genomes to

phenomes and connecting evolutional correspondences among multiple species.

OmicBrowse integrates multiple data servers into a single omic space through

secure peer-to-peer server communications, so that a user can easily obtain an

integrated view of distributed data servers, e.g. an integrated view of numerous

whole-genome tiling-array data retrieved from a user's in-house private-data

server, along with various genomic annotations from public internet servers.

OmicBrowse is especially appropriate for positional-cloning purposes. It displays

both genetic maps and genomic annotations within wide chromosomal intervals and

assists a user to select candidate genes by filtering their annotations or

associated documents against user-specified keywords or ontology terms. We also

show that an omic-space chart effectively represents schemes for integrating

multiple datasets of multiple species.AVAILABILITY: OmicBrowse is developed by

the Genome-Phenome Superbrain Project and is released as free open-source

software under the GNU General Public License at http://omicspace.riken.jp.

DOI: 10.1093/bioinformatics/btl523

PMID: 17077097 [Indexed for MEDLINE]

2770. Biopolymers. 2007 Feb 15;85(3):233-40.

Virus-PLoc: a fusion classifier for predicting the subcellular localization of

viral proteins within host and virus-infected cells.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, 1954 Hua-Shan Road, Shanghai 200030, China.

Viruses can reproduce their progenies only within a host cell, and their actions

depend both on its destructive tendencies toward a specific host cell and on

environmental conditions. Therefore, knowledge of the subcellular localization of

viral proteins in a host cell or virus-infected cell is very useful for in-depth

studying of their functions and mechanisms as well as designing antiviral drugs.

An analysis on the Swiss-Prot database (version 50.0, released on May 30, 2006)

indicates that only 23.5% of viral protein entries are annotated for their

subcellular locations in this regard. As for the gene ontology database, the

corresponding percentage is 23.8%. Such a gap calls for the development of high

throughput tools for timely annotating the localization of viral proteins within

host and virus-infected cells. In this article, a predictor called "Virus-PLoc"

has been developed that is featured by fusing many basic classifiers with each

engineered according to the K-nearest neighbor rule. The overall jackknife

success rate obtained by Virus-PLoc in identifying the subcellular compartments

of viral proteins was 80% for a benchmark dataset in which none of proteins has

more than 25% sequence identity to any other in a same location site. Virus-PLoc

will be freely available as a web-server at http://202.120.37.186/bioinf/virus

for the public usage. Furthermore, Virus-PLoc has been used to provide

large-scale predictions of all viral protein entries in Swiss-Prot database that

do not have subcellular location annotations or are annotated as being uncertain.

The results thus obtained have been deposited in a downloadable file prepared

with Microsoft Excel and named "Tab\_Virus-PLoc.xls." This file is available at

the same website and will be updated twice a year to include the new entries of

viral proteins and reflect the continuous development of Virus-PLoc.

2006 Wiley Periodicals, Inc.

DOI: 10.1002/bip.20640

PMID: 17120237 [Indexed for MEDLINE]

2771. J Cell Biochem. 2007 Feb 15;100(3):665-78.

Large-scale plant protein subcellular location prediction.

Chou KC(1), Shen HB.

Author information:

(1)Gordon Life Science Institute, San Diego, CA 92130, USA. kchou@san.rr.com

Current plant genome sequencing projects have called for development of novel and

powerful high throughput tools for timely annotating the subcellular location of

uncharacterized plant proteins. In view of this, an ensemble classifier,

Plant-PLoc, formed by fusing many basic individual classifiers, has been

developed for large-scale subcellular location prediction for plant proteins.

Each of the basic classifiers was engineered by the K-Nearest Neighbor (KNN)

rule. Plant-PLoc discriminates plant proteins among the following 11 subcellular

locations: (1) cell wall, (2) chloroplast, (3) cytoplasm, (4) endoplasmic

reticulum, (5) extracell, (6) mitochondrion, (7) nucleus, (8) peroxisome, (9)

plasma membrane, (10) plastid, and (11) vacuole. As a demonstration, predictions

were performed on a stringent benchmark dataset in which none of the proteins

included has > or =25% sequence identity to any other in a same subcellular

location to avoid the homology bias. The overall success rate thus obtained was

32-51% higher than the rates obtained by the previous methods on the same

benchmark dataset. The essence of Plant-PLoc in enhancing the prediction quality

and its significance in biological applications are discussed. Plant-PLoc is

accessible to public as a free web-server at:

(http://202.120.37.186/bioinf/plant). Furthermore, for public convenience,

results predicted by Plant-PLoc have been provided in a downloadable file at the

same website for all plant protein entries in the Swiss-Prot database that do not

have subcellular location annotations, or are annotated as being uncertain. The

large-scale results will be updated twice a year to include new entries of plant

proteins and reflect the continuous development of Plant-PLoc.

DOI: 10.1002/jcb.21096

PMID: 16983686 [Indexed for MEDLINE]

2772. Proteins. 2007 Feb 15;66(3):664-70.

MUPRED: a tool for bridging the gap between template based methods and sequence

profile based methods for protein secondary structure prediction.

Bondugula R(1), Xu D.

Author information:

(1)Digital Biology Laboratory, Department of Computer Science, Christopher S.

Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA.

Predicting secondary structures from a protein sequence is an important step for

characterizing the structural properties of a protein. Existing methods for

protein secondary structure prediction can be broadly classified into template

based or sequence profile based methods. We propose a novel framework that

bridges the gap between the two fundamentally different approaches. Our framework

integrates the information from the fuzzy k-nearest neighbor algorithm and

position-specific scoring matrices using a neural network. It combines the

strengths of the two methods and has a better potential to use the information in

both the sequence and structure databases than existing methods. We implemented

the framework into a software system MUPRED. MUPRED has achieved three-state

prediction accuracy (Q3) ranging from 79.2 to 80.14%, depending on which

benchmark dataset is used. A higher Q3 can be achieved if a query protein has a

significant sequence identity (>25%) to a template in PDB. MUPRED also estimates

the prediction accuracy at the individual residue level more quantitatively than

existing methods. The MUPRED web server and executables are freely available at

http://digbio.missouri.edu/mupred.

2006 Wiley-Liss, Inc.

DOI: 10.1002/prot.21177

PMID: 17109407 [Indexed for MEDLINE]

2773. Cancer Inform. 2007 Feb 3;3:1-9.

Asterias: a parallelized web-based suite for the analysis of expression and aCGH

data.

Alibés A(1), Morrissey ER, Cañada A, Rueda OM, Casado D, Yankilevich P,

Díaz-Uriarte R.

Author information:

(1)Statistical Computing Team, Structural and Computational Biology Programme,

Spanish National Cancer Center (CNIO), Melchor Fernández Almagro 3, Madrid,

Spain.

The analysis of expression and CGH arrays plays a central role in the study of

complex diseases, especially cancer, including finding markers for early

diagnosis and prognosis, choosing an optimal therapy, or increasing our

understanding of cancer development and metastasis. Asterias

(http://www.asterias.info) is an integrated collection of freely-accessible web

tools for the analysis of gene expression and aCGH data. Most of the tools use

parallel computing (via MPI) and run on a server with 60 CPUs for computation;

compared to a desktop or server-based but not parallelized application,

parallelization provides speed ups of factors up to 50. Most of our applications

allow the user to obtain additional information for user-selected genes

(chromosomal location, PubMed ids, Gene Ontology terms, etc.) by using clickable

links in tables and/or figures. Our tools include: normalization of expression

and aCGH data (DNMAD); converting between different types of gene/clone and

protein identifiers (IDconverter/IDClight); filtering and imputation (preP);

finding differentially expressed genes related to patient class and survival data

(Pomelo II); searching for models of class prediction (Tnasas); using random

forests to search for minimal models for class prediction or for large subsets of

genes with predictive capacity (GeneSrF); searching for molecular signatures and

predictive genes with survival data (SignS); detecting regions of genomic DNA

gain or loss (ADaCGH). The capability to send results between different

applications, access to additional functional information, and parallelized

computation make our suite unique and exploit features only available to

web-based applications.

PMCID: PMC2675829

PMID: 19455230

2774. Bioinformatics. 2007 Feb 1;23(3):383-4. Epub 2007 Jan 3.

SChiSM2: creating interactive web page annotations of molecular structure models

using Jmol.

Cammer S(1).

Author information:

(1)Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State

University, Blacksburg, VA, USA. cammer@vbi.vt.edu

SChiSM2 is a web server-based program for creating web pages that include

interactive molecular graphics using the freely-available applet, Jmol, for

illustration. The program works with Internet Explorer and Firefox on Windows,

Safari and Firefox on Mac OSX and Firefox on Linux.AVAILABILITY: The program can

be accessed at the following address: http://ci.vbi.vt.edu/cammer/schism2.html.

DOI: 10.1093/bioinformatics/btl603

PMID: 17204464 [Indexed for MEDLINE]

2775. Bioinformatics. 2007 Feb 1;23(3):392-3. Epub 2006 Nov 30.

BioNetBuilder: automatic integration of biological networks.

Avila-Campillo I(1), Drew K, Lin J, Reiss DJ, Bonneau R.

Author information:

(1)Department of Biology, New York University, New York, NY, USA.

iliana\_avila-campillo@merck.com

BioNetBuilder is an open-source client-server Cytoscape plugin that offers a

user-friendly interface to create biological networks integrated from several

databases. Users can create networks for approximately 1500 organisms, including

common model organisms and human. Currently supported databases include: DIP,

BIND, Prolinks, KEGG, HPRD, The BioGrid and GO, among others. The BioNetBuilder

plugin client is available as a Java Webstart, providing a platform-independent

network interface to these public databases.AVAILABILITY:

http://err.bio.nyu.edu/cytoscape/bionetbuilder/

DOI: 10.1093/bioinformatics/btl604

PMID: 17138585 [Indexed for MEDLINE]

2776. Cyberpsychol Behav. 2007 Feb;10(1):57-63.

Special online consulting for patients with eating disorders and their relatives:

analysis of user characteristics and E-mail content.

Grunwald M(1), Wesemann D.

Author information:

(1)EEG Research Laboratory, Department of Psychiatry, University of Leipzig,

Leipzig, Germany. mgrun@medizin.uni-leipzig.de

In the treatment of chronic diseases, programs that use the internet as a medium

are becoming more and more important as a complement to classical intervention

techniques. Since 1998, a non-profit information and online consulting service

for patients with eating disorders and their friends and relatives

(www.ab-server.de) has existed. This was established by members of the Deutsche

Forschungsinitiative Essstörungen e.V. (DFE) [German Research Initiative for

Eating Disorders] and members of the Clinic of Psychiatry, University of Leipzig,

Germany. For the present study, 2,176 e-mail requests from users of the online

consultation service were analyzed qualitatively and quantitatively in order to

better understand the differences between different types and groups of users.

The analysis was related to the social field of the person requesting the

consultation, the type of disorder reported, and the content of the e-mail

request. Three main user groups could be identified: people who described

themselves as having an eating disorder (57.2%), people who were related socially

to the affected person (32.4%), and interested persons (9.8%). The consulting

service was predominantly used by persons suffering from bulimia nervosa or their

families and friends (63.1%). One third (33.3%) of the posted e-mails were

related to behavioral patterns in dealing with the illness and the affected

person. They were followed by inquiries for information about the disease (18.7%)

and by those seeking help in finding specialized clinics/therapists and places in

therapies. The increasing use of the online consulting service indicates that

there is a substantial need for information and help in persons with eating

disorders and in their relatives, who are able to easily contact professionals

using this online service. Online consulting has a high potential for

complementary care of affected people.

DOI: 10.1089/cpb.2006.9992

PMID: 17305449 [Indexed for MEDLINE]

2777. Protein Sci. 2007 Feb;16(2):239-49. Epub 2006 Dec 22.

Redesigning protein pKa values.

Tynan-Connolly BM(1), Nielsen JE.

Author information:

(1)School of Biomolecular and Biomedical Science, Centre for Synthesis and

Chemical Biology, UCD Conway Institute, University College Dublin, Belfield,

Dublin 4, Ireland.

The ability to re-engineer enzymatic pH-activity profiles is of importance for

industrial applications of enzymes. We theoretically explore the feasibility of

re-engineering enzymatic pH-activity profiles by changing active site pK(a)

values using point mutations. We calculate the maximum achievable DeltapK(a)

values for 141 target titratable groups in seven enzymes by introducing

conservative net-charge altering point mutations. We examine the importance of

the number of mutations introduced, their distance from the target titratable

group, and the characteristics of the target group itself. The results show that

multiple mutations at 10A can change pK(a) values up to two units, but that the

introduction of a requirement to keep other pK(a) values constant reduces the

magnitude of the achievable DeltapK(a). The algorithm presented shows a good

correlation with existing experimental data and is available for download and via

a web server at http://enzyme.ucd.ie/pKD.

DOI: 10.1110/ps.062538707

PMCID: PMC2203286

PMID: 17189477 [Indexed for MEDLINE]

2778. Bioinformatics. 2007 Jan 15;23(2):e205-11.

Vorolign--fast structural alignment using Voronoi contacts.

Birzele F(1), Gewehr JE, Csaba G, Zimmer R.

Author information:

(1)Practical Informatics and Bioinformatics Group, Department of Informatics,

Ludwig-Maximilians-University, Munich, Germany. fabian.birzele@ifi.lmu.de

Vorolign, a fast and flexible structural alignment method for two or more protein

structures is introduced. The method aligns protein structures using double

dynamic programming and measures the similarity of two residues based on the

evolutionary conservation of their corresponding Voronoi-contacts in the protein

structure. This similarity function allows aligning protein structures even in

cases where structural flexibilities exist. Multiple structural alignments are

generated from a set of pairwise alignments using a consistency-based,

progressive multiple alignment strategy.RESULTS: The performance of Vorolign is

evaluated for different applications of protein structure comparison, including

automatic family detection as well as pairwise and multiple structure alignment.

Vorolign accurately detects the correct family, superfamily or fold of a protein

with respect to the SCOP classification on a set of difficult target structures.

A scan against a database of >4000 proteins takes on average 1 min per target.

The performance of Vorolign in calculating pairwise and multiple alignments is

found to be comparable with other pairwise and multiple protein structure

alignment methods.

AVAILABILITY: Vorolign is freely available for academic users as a web server at

http://www.bio.ifi.lmu.de/Vorolign

DOI: 10.1093/bioinformatics/btl294

PMID: 17237093 [Indexed for MEDLINE]

2779. Bioinformatics. 2007 Jan 15;23(2):e110-5.

Phylogenetic reconstruction from non-genomic data.

Clemente JC(1), Satou K, Valiente G.

Author information:

(1)School of Knowledge Science, Japan Advanced Institute of Science and

Technology, 1-1 Asahidai, Nomi Ishikawa 923-1292, Japan. clemente@jaist.ac.jp

MOTIVATION: Recent results related to horizontal gene transfer suggest that

phylogenetic reconstruction cannot be determined conclusively from sequence data,

resulting in a shift from approaches based on polymorphism information in DNA or

protein sequence to studies aimed at understanding the evolution of complete

biological processes. The increasing amount of available information on metabolic

pathways for several species makes it of greater relevance to understand the

similarities and differences among such pathways. These similarities can then be

used to infer phylogenetic trees not based exclusively in sequence data,

therefore avoiding the previously mentioned problems.

RESULTS: In this article, we present a method to assess the structural similarity

of metabolic pathways for several organisms. Our algorithms work by using one of

the three possible enzyme similarity measures (hierarchical, information content,

gene ontology), and one of the two clustering methods (neighbor-joining,

unweighted pair group method with arithmetic mean), to produce a phylogenetic

tree both in Newick and graphic format. The web server implementing our

algorithms is optimized to answer queries in linear time.

AVAILABILITY: The software is available for free public use on a web server, at

the address http://www.jaist.ac.jp/~clemente/cgi-bin/phylo.pl. It is available on

demand in source code form for research use to educational institutions,

non-profit research institutes, government research laboratories and individuals,

for non-exclusive use, without the right of the licensee to further redistribute

the source code.

DOI: 10.1093/bioinformatics/btl307

PMID: 17237077 [Indexed for MEDLINE]

2780. Biol Direct. 2007 Jan 12;2:1.

pkaPS: prediction of protein kinase A phosphorylation sites with the simplified

kinase-substrate binding model.

Neuberger G(1), Schneider G, Eisenhaber F.

Author information:

(1)IMP - Research Institute of Molecular Pathology, Dr, Bohr-Gasse 7, A-1030

Vienna, Austria. neuberger@imp.univie.ac.at

BACKGROUND: Protein kinase A (cAMP-dependent kinase, PKA) is a serine/threonine

kinase, for which ca. 150 substrate proteins are known. Based on a refinement of

the recognition motif using the available experimental data, we wished to apply

the simplified substrate protein binding model for accurate prediction of PKA

phosphorylation sites, an approach that was previously successful for the

prediction of lipid posttranslational modifications and of the PTS1 peroxisomal

translocation signal.

RESULTS: Approximately 20 sequence positions flanking the phosphorylated residue

on both sides have been found to be restricted in their sequence variability

(region -18...+23 with the site at position 0). The conserved physical pattern

can be rationalized in terms of a qualitative binding model with the catalytic

cleft of the protein kinase A. Positions -6...+4 surrounding the phosphorylation

site are influenced by direct interaction with the kinase in a varying degree.

This sequence stretch is embedded in an intrinsically disordered region composed

preferentially of hydrophilic residues with flexible backbone and small side

chain. This knowledge has been incorporated into a simplified analytical model of

productive binding of substrate proteins with PKA.

CONCLUSION: The scoring function of the pkaPS predictor can confidently

discriminate PKA phosphorylation sites from serines/threonines with

non-permissive sequence environments (sensitivity of appoximately 96% at a

specificity of approximately 94%). The tool "pkaPS" has been applied on the whole

human proteome. Among new predicted PKA targets, there are entirely

uncharacterized protein groups as well as apparently well-known families such as

those of the ribosomal proteins L21e, L22 and L6.

AVAILABILITY: The supplementary data as well as the prediction tool as WWW server

are available at http://mendel.imp.univie.ac.at/sat/pkaPS.

REVIEWERS: Erik van Nimwegen (Biozentrum, University of Basel, Switzerland),

Sandor Pongor (International Centre for Genetic Engineering and Biotechnology,

Trieste, Italy), Igor Zhulin (University of Tennessee, Oak Ridge National

Laboratory, USA).

DOI: 10.1186/1745-6150-2-1

PMCID: PMC1783638

PMID: 17222345

2781. BMC Bioinformatics. 2007 Jan 5;8:4.

VaxiJen: a server for prediction of protective antigens, tumour antigens and

subunit vaccines.

Doytchinova IA(1), Flower DR.

Author information:

(1)Faculty of Pharmacy, Medical University of Sofia, 2 Dunav St., 1000 Sofia,

Bulgaria. idoytchinova@pharmfac.acad.bg <idoytchinova@pharmfac.acad.bg>

BACKGROUND: Vaccine development in the post-genomic era often begins with the in

silico screening of genome information, with the most probable protective

antigens being predicted rather than requiring causative microorganisms to be

grown. Despite the obvious advantages of this approach--such as speed and cost

efficiency--its success remains dependent on the accuracy of antigen prediction.

Most approaches use sequence alignment to identify antigens. This is problematic

for several reasons. Some proteins lack obvious sequence similarity, although

they may share similar structures and biological properties. The antigenicity of

a sequence may be encoded in a subtle and recondite manner not amendable to

direct identification by sequence alignment. The discovery of truly novel

antigens will be frustrated by their lack of similarity to antigens of known

provenance. To overcome the limitations of alignment-dependent methods, we

propose a new alignment-free approach for antigen prediction, which is based on

auto cross covariance (ACC) transformation of protein sequences into uniform

vectors of principal amino acid properties.

RESULTS: Bacterial, viral and tumour protein datasets were used to derive models

for prediction of whole protein antigenicity. Every set consisted of 100 known

antigens and 100 non-antigens. The derived models were tested by internal

leave-one-out cross-validation and external validation using test sets. An

additional five training sets for each class of antigens were used to test the

stability of the discrimination between antigens and non-antigens. The models

performed well in both validations showing prediction accuracy of 70% to 89%. The

models were implemented in a server, which we call VaxiJen.

CONCLUSION: VaxiJen is the first server for alignment-independent prediction of

protective antigens. It was developed to allow antigen classification solely

based on the physicochemical properties of proteins without recourse to sequence

alignment. The server can be used on its own or in combination with

alignment-based prediction methods. It is freely-available online at the URL:

http://www.jenner.ac.uk/VaxiJen.

DOI: 10.1186/1471-2105-8-4

PMCID: PMC1780059

PMID: 17207271 [Indexed for MEDLINE]

2782. BMC Bioinformatics. 2007 Jan 3;8:2.

TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein

sequences.

Karpenahalli MR(1), Lupas AN, Söding J.

Author information:

(1)Department of Protein Evolution, Max-Planck-lnstitute for Developmental

Biology, Spemannstrasse 35, D-72076 Tübingen, Germany. manju@tuebingen.mpg.de

<manju@tuebingen.mpg.de>

BACKGROUND: Solenoid repeat proteins of the Tetratrico Peptide Repeat (TPR)

family are involved as scaffolds in a broad range of protein-protein

interactions. Several resources are available for the prediction of TPRs,

however, they often fail to detect divergent repeat units.

RESULTS: We have developed TPRpred, a profile-based method which uses a

P-value-dependent score offset to include divergent repeat units and which

exploits the tendency of repeats to occur in tandem. TPRpred detects not only

TPR-like repeats, but also the related Pentatrico Peptide Repeats (PPRs) and

SEL1-like repeats. The corresponding profiles were generated through iterative

searches, by varying the threshold parameters for inclusion of repeat units into

the profiles, and the best profiles were selected based on their performance on

proteins of known structure. We benchmarked the performance of TPRpred in

detecting TPR-containing proteins and in delineating the individual repeats

therein, against currently available resources.

CONCLUSION: TPRpred performs significantly better in detecting divergent repeats

in TPR-containing proteins, and finds more individual repeats than the existing

methods. The web server is available at http://tprpred.tuebingen.mpg.de, and the

C++ and Perl sources of TPRpred along with the profiles can be downloaded from

ftp://ftp.tuebingen.mpg.de/ebio/protevo/TPRpred/.

DOI: 10.1186/1471-2105-8-2

PMCID: PMC1774580

PMID: 17199898 [Indexed for MEDLINE]

2783. Bioinformatics. 2007 Jan 1;23(1):125-6. Epub 2006 Oct 31.

GECO--linear visualization for comparative genomics.

Kuenne CT(1), Ghai R, Chakraborty T, Hain T.

Author information:

(1)Institute of Medical Microbiology, Justus-Liebig-University Frankfurter

Strasse 107, D-35392 Giessen, Germany.

In order to understand and interpret phylogenetic and functional relationships

between multiple prokaryotic species, qualitative and quantitative data must be

correlated and displayed. GECO allows linear visualization of multiple genomes

using a client/server based approach by dynamically creating .png- or

.pdf-formatted images. It is able to display ortholog relations calculated using

BLASTCLUST by color coding ortholog representations. Irregularities on the

genomic level can be identified by anomalous G/C composition. Thus, this software

will enable researchers to detect horizontally transferred genes, pseudogenes and

insertions/deletions in related microbial genomes.AVAILABILITY:

http://bioinfo.mikrobio.med.uni-giessen.de/geco2/GecoMainServlet

DOI: 10.1093/bioinformatics/btl556

PMID: 17077098 [Indexed for MEDLINE]

2784. Bioinformatics. 2007 Jan 1;23(1):127-8. Epub 2006 Oct 18.

Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and

annotation.

Letunic I(1), Bork P.

Author information:

(1)EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

Interactive Tree Of Life (iTOL) is a web-based tool for the display, manipulation

and annotation of phylogenetic trees. Trees can be interactively pruned and

re-rooted. Various types of data such as genome sizes or protein domain

repertoires can be mapped onto the tree. Export to several bitmap and vector

graphics formats is supported.AVAILABILITY: iTOL is available at

http://itol.embl.de

DOI: 10.1093/bioinformatics/btl529

PMID: 17050570 [Indexed for MEDLINE]

2785. Comput Syst Bioinformatics Conf. 2007;6:381-4.

CBioC: beyond a prototype for collaborative annotation of molecular interactions

from the literature.

Baral C(1), Gonzalez G, Gitter A, Teegarden C, Zeigler A, Joshi-Topé G.

Author information:

(1)School of Computing and Informatics, Arizona State University, Tempe, AZ

85281, USA. chitta@asu.edu

In molecular biology research, looking for information on a particular entity

such as a gene or a protein may lead to thousands of articles, making it

impossible for a researcher to individually read these articles and even just

their abstracts. Thus, there is a need to curate the literature to get various

nuggets of knowledge, such as an interaction between two proteins, and store them

in a database. However the body of existing biomedical articles is growing at a

very fast rate, making it impossible to curate them manually. An alternative

approach of using computers for automatic extraction has problem with accuracy.

We propose to leverage the advantages of both techniques, extracting binary

relationships between biological entities automatically from the biomedical

literature and providing a platform that allows community collaboration in the

annotation of the extracted relationships. Thus, the community of researchers

that writes and reads the biomedical texts can use the server for searching our

database of extracted facts, and as an easy-to-use web platform to annotate facts

relevant to them. We presented a preliminary prototype as a proof of concept

earlier(1). This paper presents the working implementation available for download

at http://www.cbioc.org as a browser-plug in for both Internet Explorer and

FireFox. This current version has been available since June of 2006, and has over

160 registered users from around the world. Aside from its use as an annotation

tool, data from CBioC has also been used in computational methods with

encouraging results.

PMID: 17951840 [Indexed for MEDLINE]

2786. Curr Protoc Bioinformatics. 2007 Jan;Chapter 1:Unit 1.15. doi:

10.1002/0471250953.bi0115s16.

Using the Ensembl genome server to browse genomic sequence data.

Fernández-Suárez XM(1), Schuster MK.

Author information:

(1)European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, United

Kingdom.

The Ensembl genome Web browser (http://www.ensembl.org) provides a comprehensive

source of automatic annotation of the human genome sequence (as well as other

species of biomedical interest), with confirmed gene predictions that have been

integrated with external data sources. This unit describes how to use the Ensembl

browser, how to find your gene or protein of interest and get information and

external links about them, and how to use the comparative genomic data.

DOI: 10.1002/0471250953.bi0115s16

PMID: 18428779 [Indexed for MEDLINE]

2787. Genome Biol. 2007;8(12):R258.

Broad network-based predictability of Saccharomyces cerevisiae gene

loss-of-function phenotypes.

McGary KL(1), Lee I, Marcotte EM.

Author information:

(1)Center for Systems and Synthetic Biology, Institute for Cellular and Molecular

Biology, University of Texas at Austin, 2500 Speedway, Austin, Texas 78712, USA.

kris.mcgary@mail.utexas.edu

We demonstrate that loss-of-function yeast phenotypes are predictable by

guilt-by-association in functional gene networks. Testing 1,102 loss-of-function

phenotypes from genome-wide assays of yeast reveals predictability of diverse

phenotypes, spanning cellular morphology, growth, metabolism, and quantitative

cell shape features. We apply the method to extend a genome-wide screen by

predicting, then verifying, genes whose disruption elongates yeast cells, and to

predict human disease genes. To facilitate network-guided screens, a web server

is available http://www.yeastnet.org.

DOI: 10.1186/gb-2007-8-12-r258

PMCID: PMC2246260

PMID: 18053250 [Indexed for MEDLINE]

2788. In Silico Biol. 2007;7(4-5):405-12.

BTXpred: prediction of bacterial toxins.

Saha S(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology Sector-39A, Chandigarh, India.

This paper describes a method developed for predicting bacterial toxins from

their amino acid sequences. All the modules, developed in this study, were

trained and tested on a non-redundant dataset of 150 bacterial toxins that

included 77 exotoxins and 73 endotoxins. Firstly, support vector machines (SVM)

based modules were developed for predicting the bacterial toxins using amino

acids and dipeptides composition and achieved an accuracy of 96.07% and 92.50%,

respectively. Secondly, SVM based modules were developed for discriminating

entotoxins and exotoxins, using amino acids and dipeptides composition and

achieved an accuracy of 95.71% and 92.86%, respectively. In addition, modules

have been developed for classifying the exotoxins (e.g. activate adenylate

cyclase, activate guanylate cyclase, neurotoxins) using hidden Markov models

(HMM), PSI-BLAST and a combination of the two and achieved overall accuracy of

95.75%, 97.87% and 100%, respectively. Based on the above study, a web server

called 'BTXpred' has been developed, which is available at

http://www.imtech.res.in/raghava/btxpred/. Supplementary information is available

at http://www.imtech.res.in/raghava/btxpred/supplementary.html.

PMID: 18391233 [Indexed for MEDLINE]

2789. In Silico Biol. 2007;7(2):145-50.

DomainDraw: a macromolecular feature drawing program.

Fink JL(1), Hamilton N.

Author information:

(1)ARC Centre in Bioinformatics, University of Queensland, St Lucia, Queensland,

Australia. l.fink@imb.uq.edu.au

Visualization of functional and structural features of biological macromolecules

is an important aspect of communicating and analyzing biological data, for

example the presence of a transmembrane domain in relation to a nucleotide

binding site or the organization of transcription factor binding sites in a

promoter. However, this is not necessarily a trivial task especially when the

feature information is complex or lengthy. While there are some tools available

that can create these images, none have been implemented for the specific purpose

of automating the generation of presentation-quality graphics for displaying

feature information. We have implemented DomainDraw, a visualization tool that

can be used to generate schematic diagrams of biological macromolecules for the

purpose of representing the relative position and range of user-specified domains

or motifs. The user specifies the name, position, and range of the domains of

interest and DomainDraw generates the image based on these parameters. Optional

parameters include domain color and shape, image size, and whether to align

multiple proteins using a particular domain. DomainDraw is publicly available as

a web server and can be accessed at http://domaindraw.imb.uq.edu.au. The

executable may be obtained by contacting the authors.

PMID: 17688439 [Indexed for MEDLINE]

2790. J Biosci. 2007 Jan;32(1):97-100.

SURF'S UP! - protein classification by surface comparisons.

Sasin JM(1), Godzik A, Bujnicki JM.

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology, Warsaw, Poland. asia@genesilico.pl

Large-scale genome sequencing and structural genomics projects generate numerous

sequences and structures for 'hypothetical' proteins without functional

characterizations. Detection of homology to experimentally characterized proteins

can provide functional clues, but the accuracy of homology-based predictions is

limited by the paucity of tools for quantitative comparison of diverging residues

responsible for the functional divergence. SURF'S UP! is a web server for

analysis of functional relationships in protein families, as inferred from

protein surface maps comparison according to the algorithm. It assigns a

numerical score to the similarity between patterns of physicochemical

features(charge, hydrophobicity) on compared protein surfaces. It allows

recognizing clusters of proteins that have similar surfaces, hence presumably

similar functions. The server takes as an input a set of protein coordinates and

returns files with "spherical coordinates" of proteins in a PDB format and their

graphical presentation, a matrix with values of mutual similarities between the

surfaces, and the unrooted tree that represents the clustering of similar

surfaces, calculated by the neighbor-joining method. SURF'S UP! facilitates the

comparative analysis of physicochemical features of the surface, which are the

key determinants of the protein function. By concentrating on coarse surface

features, SURF'S UP! can work with models obtained from comparative modelling.

Although it is designed to analyse the conservation among homologs, it can also

be used to compare surfaces of non-homologous proteins with different

three-dimensional folds, as long as a functionally meaningful structural

superposition is supplied by the user. Another valuable characteristic of our

method is the lack of initial assumptions about the functional features to be

compared. SURF'S UP! is freely available for academic researchers at

http://asia.genesilico.pl/surfs\_up/.

PMID: 17426383 [Indexed for MEDLINE]

2791. Methods Mol Biol. 2007;409:381-6. doi: 10.1007/978-1-60327-118-9\_28.

TAPPred prediction of TAP-binding peptides in antigens.

Bhasin M(1), Lata S, Raghava GP.

Author information:

(1)Institute of Microbial Technology, Chandigarh, India.

The transporter associated with antigen processing (TAP) plays a crucial role in

the transport of the peptide fragments of the proteolysed antigenic or

self-altered proteins to the endoplasmic reticulum where the association between

these peptides and the major histocompatibility complex (MHC) class I molecules

takes place. Therefore, prediction of TAP-binding peptides is highly helpful in

identifying the MHC class I-restricted T-cell epitopes and hence in the subunit

vaccine designing. In this chapter, we describe a support vector machine

(SVM)-based method TAPPred that allows users to predict TAP-binding affinity of

peptides over web. The server allows user to predict TAP binders using a simple

SVM model or cascade SVM model. The server also allows user to customize the

display/output. It is freely available for academicians and noncommercial

organization at the address http://www.imtech.res.in/raghava/tappred.

DOI: 10.1007/978-1-60327-118-9\_28

PMID: 18450016 [Indexed for MEDLINE]

2792. Methods Mol Biol. 2007;406:275-99.

BGI-RIS V2.

He X(1), Wang J.

Author information:

(1)Beijing Genomics Institute, Chinese Academy of Sciences, Beijing, China.

Rice serves as both a staple for over half of the world's population and a model

organism for plants of the grass family. Beijing Genomics Institute (BGI) has

long been engaged in rice genomic research: sequencing, assembly, information

analysis and integration. Such intensive research results in public releases and

biological applications. In order to facilitate obtaining and operating on the

rice genomic data, as well as to provide a genomic groundwork for comparative,

functional or evolutionary research on important cereal crops, BGI has

established and updated the Rice Information System (BGI-RIS V2), an integrated

information resource and comparative analysis workbench for rice genomes. BGI-RIS

V2 offers not only genomic sequences, which combine the genomic data of Oryza

sativa L. ssp. indica (by BGI) with Oryza sativa L. ssp. japonica, but also most

detailed annotation data, including genetic markers, Bacterial Artificial

Chromosome (BAC) end sequences, gene contents, cDNAs, oligos, tiling arrays,

repetitive elements, and genomic polymorphisms. As a basic platform, BGI-RIS V2

also offers graphical interfaces and a series of tools and services for gene

finding, genomic alignment and genomic assembly. This database is available

through the web server (http://rise.genomics.org.cn or

http://rice.genomics.org.cn) and the File Transfer Protocol (FTP) server

(ftp://ftp.genomics.org.cn/pub/database/rice).

PMID: 18287698 [Indexed for MEDLINE]

2793. Methods Mol Biol. 2007;396:71-91.

Gene annotation and pathway mapping in KEGG.

Aoki-Kinoshita KF, Kanehisa M.

KEGG is a database resource (http://www.genome.jp/kegg/) that provides all

knowledge about genomes and their relationships to biological systems such as

cells and whole organisms as well as their interactions with the environment.

KEGG is categorized in terms of building blocks in the genomic space, known as

KEGG GENES, the chemical space, KEGG LIGAND, as well as wiring diagrams of

interaction and reaction networks, known as KEGG PATHWAY. A fourth database

called KEGG BRITE was also recently incorporated to provide computerized

annotations and pathway reconstruction based on the current KEGG knowledgebase.

KEGG BRITE contains KEGG Orthology (KO), a classification of ortholog and paralog

groups based on highly confident sequence similarity scores, and the reaction

classification system for biochemical reaction classification, along with other

classifications for compounds and drugs. BRITE is also the basis for the KEGG

Automatic Annotation Server (KAAS), which automatically annotates a given set of

genes and correspondingly generates pathway maps. This chapter introduces KEGG

and its various tools for genomic analyses, focusing on the usage of the KEGG

GENES, PATHWAY, and BRITE resources and the KAAS tool (see Note 1).

DOI: 10.1007/978-1-59745-515-2\_6

PMID: 18025687 [Indexed for MEDLINE]

2794. Methods Mol Biol. 2007;395:255-68.

Improving pairwise sequence alignment between distantly related proteins.

Feng JA(1).

Author information:

(1)Department of Chemistry, Center for Biotechnology, Temple University, USA.

Sequence alignment between remotely related proteins has been one of the more

difficult problems in structural biology. Improvements have been achieved by

incorporating information that enhances the diversity of the substitution

matrices. NdPASA is a web-based server that optimizes sequence alignments between

proteins sharing low percentages of sequence identity. The program integrates

structure information of the template sequence into a global alignment algorithm

by employing amino acids' neighbor-dependent propensities for secondary structure

as unique parameters for alignment. NdPASA optimizes alignment by evaluating the

likelihood of a residue pair in the query sequence matching against a

corresponding residue pair adopting a particular secondary structure in the

template sequence. The server is designed to aid homologous protein structure

modeling. It is most effective when the structure of the template sequence is

known. NdPASA can be accessed online at www.fenglab.org/bioserver.html.

PMID: 17993679 [Indexed for MEDLINE]

2795. Methods Mol Biol. 2007;395:237-54.

Mulan: multiple-sequence alignment to predict functional elements in genomic

sequences.

Loots GG(1), Ovcharenko I.

Author information:

(1)Lawrence Berkeley National Laboratory, USA.

Multiple sequence alignment analysis is a powerful approach for translating the

evolutionary selective power into phylogenetic relationships to localize

functional coding and noncoding genomic elements. The tool Mulan

(http://mulan.dcode.org/) has been designed to effectively perform multiple

comparisons of genomic sequences necessary to facilitate bioinformatic-driven

biological discoveries. The Mulan network server is capable of comparing both

closely and distantly related genomes to identify conserved elements over a broad

range of evolutionary time. Several novel algorithms are brought together in this

tool: the tba multisequence aligner program used to rapidly identify local

sequence conservation and the multiTF program to detect evolutionarily conserved

transcription factor binding sites in alignments. Mulan is integrated with the

ERC Browser, the UCSC Genome Browser for quick uploads of available sequences and

supports two-way communication with the GALA database to overlay GALA functional

genome annotation with sequence conservation profiles. Local multiple alignments

computed by Mulan ensure reliable representation of short- and large-scale

genomic rearrangements in distant organisms. Recently, we have also introduced

the ability to handle duplications to permit the reliable reconstruction of

evolutionary events that underlie the genome sequence data. Here, we describe the

main features of the Mulan tool that include the interactive modification of

critical conservation parameters, visualization options, and dynamic access to

sequence data from visual graphs for flexible and easy-to-perform analysis of

differentially evolving genomic regions.

PMCID: PMC3704129

PMID: 17993678 [Indexed for MEDLINE]

2796. Methods Mol Biol. 2007;395:195-204.

Alignment of genomic sequences using DIALIGN.

Morgenstern B(1).

Author information:

(1)Institute of Microbiology & Genetics, University of Göttingen.

DIALIGN is a software program for multiple alignment of DNA or protein sequences

that combines global and local alignment features. During the last years, the

program has been used extensively to compare syntenic regions in genomic

sequences. An anchoring option speeds up the alignment procedure and makes it

possible to use user-defined constraints to improve the quality of the program

output. This chapter explains features of DIALIGN that are useful if genomic

sequences are to be aligned. The program is online available through Göttingen

Bioinformatics Compute Server at http://dialign.gobics.de/.

PMID: 17993675 [Indexed for MEDLINE]

2797. Methods Mol Biol. 2007;395:17-34.

Comparative genomic analysis using the UCSC genome browser.

Karolchik D(1), Bejerano G, Hinrichs AS, Kuhn RM, Miller W, Rosenbloom KR, Zweig

AS, Haussler D, Kent WJ.

Author information:

(1)UCSC Genome Bioinformatics Group, Center for Biomolecular Science and

Engineering, University of California, Santa Cruz, CA, USA.

Comparative analysis of DNA sequence from multiple species can provide insights

into the function and evolutionary processes that shape genomes. The University

of California Santa Cruz (UCSC) Genome Bioinformatics group has developed several

tools and methodologies in its study of comparative genomics, many of which have

been incorporated into the UCSC Genome Browser (http://genome.ucsc.edu), an

easy-to-use online tool for browsing genomic data and aligned annotation "tracks"

in a single window. The comparative genomics annotations in the browser include

pairwise alignments, which aid in the identification of orthologous regions

between species, and conservation tracks that show measures of evolutionary

conservation among sets of multiply aligned species, highlighting regions of the

genome that may be functionally important. A related tool, the UCSC Table

Browser, provides a simple interface for querying, analyzing, and downloading the

data underlying the Genome Browser annotation tracks. Here, we describe a

procedure for examining a genomic region of interest in the Genome Browser,

analyzing characteristics of the region, filtering the data, and downloading data

sets for further study.

PMID: 17993665 [Indexed for MEDLINE]

2798. Methods Mol Biol. 2007;402:385-402.

BiSearch: ePCR tool for native or bisulfite-treated genomic template.

Arányi T(1), Tusnády GE.

Author information:

(1)Institute of Enzymology, BRC, Hungarian Academy of Sciences, Karolina,

Hungary.

The design of adequate primers for polymerase chain reaction (PCR) is sometimes a

difficult task. This is the case when either the target sequence harbors unusual

nucleotide motifs or the template is complex. Unusual nucleotide motifs can be

repeat elements, whereas complex templates are targets for mispriming and

alternative amplification products. Such examples are GC-rich native or

bisulfite-treated genomic DNA sequences. Bisulfite treatment leads to the

specific conversion of non-methylated cytosines to uracyls. This is the key step

of bisulfite genomic sequencing, widely used to determine DNA methylation of a

sequence. Here, we describe BiSearch Web server (http://bisearch.enzim.hu), a

primer design software created for designing primers to amplify such target

sequences. Furthermore, we developed a unique post-design primer analysis module,

to carry out genome wide searches to identify genomic mispriming sites and to

test by electronic (in silico) PCR (ePCR) for alternative PCR products. This

option is currently available on four native or bisulfite-treated mammalian

genomes.

DOI: 10.1007/978-1-59745-528-2\_20

PMID: 17951807 [Indexed for MEDLINE]

2799. Methods Mol Biol. 2007;373:25-38.

Web-based primer design software for genome-scale genotyping by pyrosequencing.

Ringquist S(1), Pecoraro C, Lu Y, Styche A, Rudert WA, Benos PV, Trucco M.

Author information:

(1)Division of Immunogenetics, Department of Pediatrics, Rangos Research Center,

Children's Hospital of Pittsburgh, University of Pittsburgh School of Medicine,

Pittsburgh, PA, USA.

Design of locus-specific primers for use during genetic analysis requires

combining information from multiple sources and can be a time-consuming process

when validating large numbers of assays. Data warehousing of genomic DNA

sequences and genetic variations when coupled with software applications for

optimizing the generation of locus-specific primers can increase the efficiency

of assay development. Selection of oligonucleotide primers for PCR and

Pyrosequencing (SOP3) software allows user-directed queries of warehoused data

collected from the human and mouse genome sequencing projects. The software

automates collection of DNA sequence flanking single-nucleotide polymorphisms

(SNPs) as well as the incorporation of locus-associated functional information,

such as whether the SNP occurs in an exon, intron, or untranslated region. SOP3

software accepts three types of user-directed input consisting of gene locus

symbols, SNP reference sequence numbers, or chromosomal physical location. For

human polymorphisms, SOP3 incorporates haplotype, ethnicity, and SNP validation

attributes. The output is a list of oligonucleotide primers recommended for

Pyrosequencing-based typing of genetic variations. SOP3 is available at the

Division of Immunogenetics computational server found at

http://imgen.ccbb.pitt.edu.

DOI: 10.1385/1-59745-377-3:25

PMID: 17185755 [Indexed for MEDLINE]

2800. Methods Mol Biol. 2007;357:297-305.

Proteomic analysis of foam cells.

Yang PY(1), Rui YC, Yang PY, Yu YL.

Author information:

(1)Department of Pharmacology, Secondary Military Medical University, Shanghai,

China.

Foam cells are characteristic pathological cells in the lesions of

atherosclerosis. Previous works have established macrophage-derived foam cell

model to study the central role of the foam cells, and analyzed the protein

expression profiles in foam cells. The reported in vitro foam cell model was

established by incubating the human U937 cells with oxidized low-density

lipoprotein. The global changes in protein expressions between U937 foam cell and

normal U937 cells were measured with two-dimensional gel electrophoresis, and

some interested proteins were tryptic-digested and then identified via mass

spectrometry after capillary liquid chromatography separation. Some of the

identified proteins were validated via the Internet links to the U937 proteomic

map provided from the Expasy Proteomics server (http://us.expasy.org). The

experimental data can provide potential markers during the inflammatory reactions

for atherosclerotic studies.

DOI: 10.1385/1-59745-214-9:297

PMID: 17172695 [Indexed for MEDLINE]

2801. Mol Diagn Ther. 2007;11(1):15-9.

The BiolAD-DB system : an informatics system for clinical and genetic data.

Nielsen DA(1), Leidner M, Haynes C, Krauthammer M, Kreek MJ.

Author information:

(1)Laboratory of the Biology of Addictive Diseases, The Rockefeller University,

New York, NY 10021, USA. nielsen@rockefeller.edu

The Biology of Addictive Diseases-Database (BiolAD-DB) system is a research

bioinformatics system for archiving, analyzing, and processing of complex

clinical and genetic data. The database schema employs design principles for

handling complex clinical information, such as response items in genetic

questionnaires. Data access and validation is provided by the BiolAD-DB client

application, which features a data validation engine tightly coupled to a

graphical user interface. Data integrity is provided by the password-protected

BiolAD-DB SQL compliant server and database. BiolAD-DB tools further provide

functionalities for generating customized reports and views. The BiolAD-DB system

schema, client, and installation instructions are freely available at

http://www.rockefeller.edu/biolad-db/.

PMID: 17286447 [Indexed for MEDLINE]

2802. Mol Syst Biol. 2007;3:114. Epub 2007 Jun 5.

The environmental fate of organic pollutants through the global microbial

metabolism.

Gómez MJ(1), Pazos F, Guijarro FJ, de Lorenzo V, Valencia A.

Author information:

(1)Centro de Astrobiología (INTA-CSIC), Ctra. Torrejón Ajalvir, Km 4. Torrejón de

Ardoz, Madrid, Spain.

The production of new chemicals for industrial or therapeutic applications

exceeds our ability to generate experimental data on their biological fate once

they are released into the environment. Typically, mixtures of organic pollutants

are freed into a variety of sites inhabited by diverse microorganisms, which

structure complex multispecies metabolic networks. A machine learning approach

has been instrumental to expose a correlation between the frequency of 149 atomic

triads (chemotopes) common in organo-chemical compounds and the global capacity

of microorganisms to metabolise them. Depending on the type of environmental fate

defined, the system can correctly predict the biodegradative outcome for 73-87%

of compounds. This system is available to the community as a web server

(http://www.pdg.cnb.uam.es/BDPSERVER). The application of this predictive tool to

chemical species released into the environment provides an early instrument for

tentatively classifying the compounds as biodegradable or recalcitrant. Automated

surveys of lists of industrial chemicals currently employed in large quantities

revealed that herbicides are the group of functional molecules more difficult to

recycle into the biosphere through the inclusive microbial metabolism.

DOI: 10.1038/msb4100156

PMCID: PMC1911198

PMID: 17551509 [Indexed for MEDLINE]

2803. Nucleic Acids Res. 2007;35(22):e150. Epub 2007 Nov 26.

STRALCP--structure alignment-based clustering of proteins.

Zemla A(1), Geisbrecht B, Smith J, Lam M, Kirkpatrick B, Wagner M, Slezak T, Zhou

CE.

Author information:

(1)Computing Applications and Research, Lawrence Livermore National Laboratory,

Livermore, CA 94550, USA. adamz@llnl.gov

Protein structural annotation and classification is an important and challenging

problem in bioinformatics. Research towards analysis of sequence-structure

correspondences is critical for better understanding of a protein's structure,

function, and its interaction with other molecules. Clustering of protein domains

based on their structural similarities provides valuable information for protein

classification schemes. In this article, we attempt to determine whether

structure information alone is sufficient to adequately classify protein

structures. We present an algorithm that identifies regions of structural

similarity within a given set of protein structures, and uses those regions for

clustering. In our approach, called STRALCP (STRucture ALignment-based Clustering

of Proteins), we generate detailed information about global and local

similarities between pairs of protein structures, identify fragments (spans) that

are structurally conserved among proteins, and use these spans to group the

structures accordingly. We also provide a web server at

http://as2ts.llnl.gov/AS2TS/STRALCP/ for selecting protein structures,

calculating structurally conserved regions and performing automated clustering.

DOI: 10.1093/nar/gkm1049

PMCID: PMC2190701

PMID: 18039711 [Indexed for MEDLINE]

2804. Nucleic Acids Res. 2007;35(10):3375-82. Epub 2007 May 3.

LOMETS: a local meta-threading-server for protein structure prediction.

Wu S(1), Zhang Y.

Author information:

(1)Center for Bioinformatics and Department of Molecular Bioscience, University

of Kansas, 2030 Becker Dr, Lawrence, KS 6604, USA.

We developed LOMETS, a local threading meta-server, for quick and automated

predictions of protein tertiary structures and spatial constraints. Nine

state-of-the-art threading programs are installed and run in a local computer

cluster, which ensure the quick generation of initial threading alignments

compared with traditional remote-server-based meta-servers. Consensus models are

generated from the top predictions of the component-threading servers, which are

at least 7% more accurate than the best individual servers based on TM-score at a

t-test significance level of 0.1%. Moreover, side-chain and C-alpha (C(alpha))

contacts of 42 and 61% accuracy respectively, as well as long- and short-range

distant maps, are automatically constructed from the threading alignments. These

data can be easily used as constraints to guide the ab initio procedures such as

TASSER for further protein tertiary structure modeling. The LOMETS server is

freely available to the academic community at

http://zhang.bioinformatics.ku.edu/LOMETS.

DOI: 10.1093/nar/gkm251

PMCID: PMC1904280

PMID: 17478507 [Indexed for MEDLINE]

2805. Nucleic Acids Res. 2007;35(2):433-40. Epub 2006 Dec 14.

More complete gene silencing by fewer siRNAs: transparent optimized design and

biophysical signature.

Ladunga I(1).

Author information:

(1)Center for Biotechnology and Department of Statistics, University of

Nebraska-Lincoln, Lincoln, NE 68588-0665, USA. sladunga@unl.edu

Highly accurate knockdown functional analyses based on RNA interference (RNAi)

require the possible most complete hydrolysis of the targeted mRNA while avoiding

the degradation of untargeted genes (off-target effects). This in turn requires

significant improvements to target selection for two reasons. First, the average

silencing activity of randomly selected siRNAs is as low as 62%. Second, applying

more than five different siRNAs may lead to saturation of the RNA-induced

silencing complex (RISC) and to the degradation of untargeted genes. Therefore,

selecting a small number of highly active siRNAs is critical for maximizing

knockdown and minimizing off-target effects. To satisfy these needs, a publicly

available and transparent machine learning tool is presented that ranks all

possible siRNAs for each targeted gene. Support vector machines (SVMs) with

polynomial kernels and constrained optimization models select and utilize the

most predictive effective combinations from 572 sequence, thermodynamic,

accessibility and self-hairpin features over 2200 published siRNAs. This tool

reaches an accuracy of 92.3% in cross-validation experiments. We fully present

the underlying biophysical signature that involves free energy, accessibility and

dinucleotide characteristics. We show that while complete silencing is possible

at certain structured target sites, accessibility information improves the

prediction of the 90% active siRNA target sites. Fast siRNA activity predictions

can be performed on our web server at http://optirna.unl.edu/.

DOI: 10.1093/nar/gkl1065

PMCID: PMC1802606

PMID: 17169992 [Indexed for MEDLINE]

2806. Nucleic Acids Res. 2007 Jan;35(Database issue):D354-7. Epub 2006 Dec 5.

PEDANT genome database: 10 years online.

Riley ML(1), Schmidt T, Artamonova II, Wagner C, Volz A, Heumann K, Mewes HW,

Frishman D.

Author information:

(1)Institute for Bioinformatics, GSF-National Research Center for Health and

Environment, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany.

The PEDANT genome database provides exhaustive annotation of 468 genomes by a

broad set of bioinformatics algorithms. We describe recent developments of the

PEDANT Web server. The all-new Graphical User Interface (GUI) implemented in

Javatrade mark allows for more efficient navigation of the genome data, extended

search capabilities, user customization and export facilities. The DNA and

Protein viewers have been made highly dynamic and customizable. We also provide

Web Services to access the entire body of PEDANT data programmatically. Finally,

we report on the application of association rule mining for automatic detection

of potential annotation errors. PEDANT is freely accessible to academic users at

http://pedant.gsf.de.

DOI: 10.1093/nar/gkl1005

PMCID: PMC1761421

PMID: 17148486 [Indexed for MEDLINE]

2807. Nucleic Acids Res. 2007 Jan;35(Database issue):D572-4. Epub 2006 Nov 29.

MINT: the Molecular INTeraction database.

Chatr-aryamontri A(1), Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli

L, Cesareni G.

Author information:

(1)Department of Biology, University of Rome Tor Vergata, Via della Ricerca

Scientifica, 00133 Rome, Italy.

The Molecular INTeraction database (MINT, http://mint.bio.uniroma2.it/mint/) aims

at storing, in a structured format, information about molecular interactions

(MIs) by extracting experimental details from work published in peer-reviewed

journals. At present the MINT team focuses the curation work on physical

interactions between proteins. Genetic or computationally inferred interactions

are not included in the database. Over the past four years MINT has undergone

extensive revision. The new version of MINT is based on a completely remodeled

database structure, which offers more efficient data exploration and analysis,

and is characterized by entries with a richer annotation. Over the past few years

the number of curated physical interactions has soared to over 95 000. The whole

dataset can be freely accessed online in both interactive and batch modes through

web-based interfaces and an FTP server. MINT now includes, as an integrated

addition, HomoMINT, a database of interactions between human proteins inferred

from experiments with ortholog proteins in model organisms

(http://mint.bio.uniroma2.it/mint/).

DOI: 10.1093/nar/gkl950

PMCID: PMC1751541

PMID: 17135203 [Indexed for MEDLINE]

2808. Nucleic Acids Res. 2007 Jan;35(Database issue):D707-10. Epub 2006 Nov 29.

Snap: an integrated SNP annotation platform.

Li S(1), Ma L, Li H, Vang S, Hu Y, Bolund L, Wang J.

Author information:

(1)The Institute of Human Genetics, University of Aarhus, DK-8000 Aarhus C,

Denmark.

Snap (Single Nucleotide Polymorphism Annotation Platform) is a server designed to

comprehensively analyze single genes and relationships between genes basing on

SNPs in the human genome. The aim of the platform is to facilitate the study of

SNP finding and analysis within the framework of medical research. Using a

user-friendly web interface, genes can be searched by name, description,

position, SNP ID or clone name. Several public databases are integrated,

including gene information from Ensembl, protein features from

Uniprot/SWISS-PROT, Pfam and DAS-CBS. Gene relationships are fetched from BIND,

MINT, KEGG and are integrated with ortholog data from TreeFam to extend the

current interaction networks. Integrated tools for primer-design and mis-splicing

analysis have been developed to facilitate experimental analysis of individual

genes with focus on their variation. Snap is available at

http://snap.humgen.au.dk/ and at http://snap.genomics.org.cn/.

DOI: 10.1093/nar/gkl969

PMCID: PMC1751554

PMID: 17135198 [Indexed for MEDLINE]

2809. Nucleic Acids Res. 2007 Jan;35(Database issue):D690-5. Epub 2006 Nov 28.

FINDbase: a relational database recording frequencies of genetic defects leading

to inherited disorders worldwide.

van Baal S(1), Kaimakis P, Phommarinh M, Koumbi D, Cuppens H, Riccardino F, Macek

M Jr, Scriver CR, Patrinos GP.

Author information:

(1)Erasmus MC, MGC-Department of Cell Biology and Genetics, Rotterdam, The

Netherlands.

Frequency of INherited Disorders database (FINDbase) (http://www.findbase.org) is

a relational database, derived from the ETHNOS software, recording frequencies of

causative mutations leading to inherited disorders worldwide. Database records

include the population and ethnic group, the disorder name and the related gene,

accompanied by links to any corresponding locus-specific mutation database, to

the respective Online Mendelian Inheritance in Man entries and the mutation

together with its frequency in that population. The initial information is

derived from the published literature, locus-specific databases and genetic

disease consortia. FINDbase offers a user-friendly query interface, providing

instant access to the list and frequencies of the different mutations. Query

outputs can be either in a table or graphical format, accompanied by reference(s)

on the data source. Registered users from three different groups, namely

administrator, national coordinator and curator, are responsible for database

curation and/or data entry/correction online via a password-protected interface.

Databaseaccess is free of charge and there are no registration requirements for

data querying. FINDbase provides a simple, web-based system for population-based

mutation data collection and retrieval and can serve not only as a valuable

online tool for molecular genetic testing of inherited disorders but also as a

non-profit model for sustainable database funding, in the form of a

'database-journal'.

DOI: 10.1093/nar/gkl934

PMCID: PMC1747180

PMID: 17135191 [Indexed for MEDLINE]

2810. Nucleic Acids Res. 2007 Jan;35(Database issue):D780-5. Epub 2006 Nov 15.

The Online Bioinformatics Resources Collection at the University of Pittsburgh

Health Sciences Library System--a one-stop gateway to online bioinformatics

databases and software tools.

Chen YB(1), Chattopadhyay A, Bergen P, Gadd C, Tannery N.

Author information:

(1)Health Sciences Library System, University of Pittsburgh, 200 Scaife Hall,

3550 Terrace Street, Pittsburgh, PA 15261, USA.

To bridge the gap between the rising information needs of biological and medical

researchers and the rapidly growing number of online bioinformatics resources, we

have created the Online Bioinformatics Resources Collection (OBRC) at the Health

Sciences Library System (HSLS) at the University of Pittsburgh. The OBRC,

containing 1542 major online bioinformatics databases and software tools, was

constructed using the HSLS content management system built on the Zope Web

application server. To enhance the output of search results, we further

implemented the Vivísimo Clustering Engine, which automatically organizes the

search results into categories created dynamically based on the textual

information of the retrieved records. As the largest online collection of its

kind and the only one with advanced search results clustering, OBRC is aimed at

becoming a one-stop guided information gateway to the major bioinformatics

databases and software tools on the Web. OBRC is available at the University of

Pittsburgh's HSLS Web site (http://www.hsls.pitt.edu/guides/genetics/obrc).

DOI: 10.1093/nar/gkl781

PMCID: PMC1669712

PMID: 17108360 [Indexed for MEDLINE]

2811. Nucleic Acids Res. 2007 Jan;35(Database issue):D47-50. Epub 2006 Nov 3.

Patome: a database server for biological sequence annotation and analysis in

issued patents and published patent applications.

Lee B(1), Kim T, Kim SK, Lee KH, Lee D.

Author information:

(1)Korean BioInformation Center, KRIBB, Daejeon 305-806, Korea.

With the advent of automated and high-throughput techniques, the number of patent

applications containing biological sequences has been increasing rapidly.

However, they have attracted relatively little attention compared to other

sequence resources. We have built a database server called Patome, which contains

biological sequence data disclosed in patents and published applications, as well

as their analysis information. The analysis is divided into two steps. The first

is an annotation step in which the disclosed sequences were annotated with RefSeq

database. The second is an association step where the sequences were linked to

Entrez Gene, OMIM and GO databases, and their results were saved as a gene-patent

table. From the analysis, we found that 55% of human genes were associated with

patenting. The gene-patent table can be used to identify whether a particular

gene or disease is related to patenting. Patome is available at

http://www.patome.org/; the information is updated bimonthly.

DOI: 10.1093/nar/gkl807

PMCID: PMC1781150

PMID: 17085479 [Indexed for MEDLINE]

2812. Nucleic Acids Res. 2007 Jan;35(Database issue):D317-21. Epub 2006 Oct 25.

TOPOFIT-DB, a database of protein structural alignments based on the TOPOFIT

method.

Leslin CM(1), Abyzov A, Ilyin VA.

Author information:

(1)Department of Biology, Northeastern University, 360 Huntington Avenue, Boston,

MA 02115, USA.

TOPOFIT-DB (T-DB) is a public web-based database of protein structural alignments

based on the TOPOFIT method, providing a comprehensive resource for comparative

analysis of protein structure families. The TOPOFIT method is based on the

discovery of a saturation point on the alignment curve (topomax point) which

presents an ability to objectively identify a border between common and variable

parts in a protein structural family, providing additional insight into protein

comparison and functional annotation. TOPOFIT also effectively detects

non-sequential relations between protein structures. T-DB provides users with the

convenient ability to retrieve and analyze structural neighbors for a protein; do

one-to-all calculation of a user provided structure against the entire current

PDB release with T-Server, and pair-wise comparison using the TOPOFIT method

through the T-Pair web page. All outputs are reported in various web-based tables

and graphics, with automated viewing of the structure-sequence alignments in the

Friend software package for complete, detailed analysis. T-DB presents

researchers with the opportunity for comprehensive studies of the variability in

proteins and is publicly available at

http://mozart.bio.neu.edu/topofit/index.php.

DOI: 10.1093/nar/gkl809

PMCID: PMC1635338

PMID: 17065464 [Indexed for MEDLINE]

2813. Pac Symp Biocomput. 2007:145-56.

BioSpider: a web server for automating metabolome annotations.

Knox C(1), Shrivastava S, Stothard P, Eisner R, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB T6G-2E8

Canada.

One of the growing challenges in life science research lies in finding useful,

descriptive or quantitative data about newly reported biomolecules (genes,

proteins, metabolites and drugs). An even greater challenge is finding

information that connects these genes, proteins, drugs or metabolites to each

other. Much of this information is scattered through hundreds of different

databases, abstracts or books and almost none of it is particularly well

integrated. While some efforts are being undertaken at the NCBI and EBI to

integrate many different databases together, this still falls short of the goal

of having some kind of human-readable synopsis that summarizes the state of

knowledge about a given biomolecule - especially small molecules. To address this

shortfall, we have developed BioSpider. BioSpider is essentially an automated

report generator designed specifically to tabulate and summarize data on

biomolecules - both large and small. Specifically, BioSpider allows users to type

in almost any kind of biological or chemical identifier (protein/gene name,

sequence, accession number, chemical name, brand name, SMILES string, InCHI

string, CAS number, etc.) and it returns an in-depth synoptic report

(approximately 3-30 pages in length) about that biomolecule and any other

biomolecule it may target. This summary includes physico-chemical parameters,

images, models, data files, descriptions and predictions concerning the query

molecule. BioSpider uses a web-crawler to scan through dozens of public databases

and employs a variety of specially developed text mining tools and locally

developed prediction tools to find, extract and assemble data for its reports.

Because of its breadth, depth and comprehensiveness, we believe BioSpider will

prove to be a particularly valuable tool for researchers in metabolomics.

BioSpider is available at: www.biospider.ca

PMID: 17990488 [Indexed for MEDLINE]

2814. Protein Eng Des Sel. 2007 Jan;20(1):39-46. Epub 2007 Jan 23.

Gpos-PLoc: an ensemble classifier for predicting subcellular localization of

Gram-positive bacterial proteins.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, 1954 Hua-Shan Road, Shanghai 200030, China.

A statistical analysis indicated that, of the 35,016 Gram-positive bacterial

proteins from the recent Swiss-Prot database, approximately 57% of these entries

are without subcellular location annotations. In the gene ontology database, the

corresponding percentage is approximately 67%, meaning the percentage of proteins

without subcellular component annotations is even higher. With the avalanche of

gene products generated in the post-genomic era, the number of such

location-unknown entries will continuously increase. It is highly desired to

develop an automated method for timely and accurately identifying their

subcellular localization because the information thus obtained is very useful for

both basic research and drug discovery practice. In view of this, an ensemble

classifier called 'Gpos-PLoc' was developed for predicting Gram-positive protein

subcellular localization. The new predictor is featured by fusing many basic

classifiers, each of which was engineered according to the optimized

evidence-theoretic K-nearest neighbors rule. As a demonstration, tests were

performed on Gram-positive proteins among the following five subcellular location

sites: (1) cell wall, (2) cytoplasm, (3) extracell, (4) periplasm and (5) plasma

membrane. To eliminate redundancy and homology bias, only those proteins which

have < 25% sequence identity to any other in a same subcellular location were

allowed to be included in the benchmark datasets. The overall success rates thus

achieved by Gpos-PLoc were > 80% for both jackknife cross-validation test and

independent dataset test, implying that Gpos-PLoc might become a very useful

vehicle for expediting the analysis of Gram-positive bacterial proteins.

Gpos-PLoc is freely accessible to public as a web-server at

http://202.120.37.186/bioinf/Gpos/. To support the need of many investigators in

the relevant areas, a downloadable file is provided at the same website to list

the results identified by Gpos-PLoc for 31,898 Gram-positive bacterial protein

entries in Swiss-Prot database that either have no subcellular location

annotation or are annotated with uncertain terms such as 'probable', 'potential',

'perhaps' and 'by similarity'. Such large-scale results will be updated once a

year to include the new entries of Gram-positive bacterial proteins and reflect

the continuous development of Gpos-PLoc.

DOI: 10.1093/protein/gzl053

PMID: 17244638 [Indexed for MEDLINE]

2815. Protein Pept Lett. 2007;14(7):669-71.

Ramachandran plot on the web (2.0).

Gopalakrishnan K(1), Sowmiya G, Sheik SS, Sekar K.

Author information:

(1)Bioinformatics Centre (Centre of Excellence in Structural Biol. &

Bio-comput.), Indian Institute of Science, Bangalore 560 012, India.

The Ramachandran plot displays the main chain conformation angles (Phi and Psi)

of the polypeptide chain of a protein molecule. The paper reports the updated

version of the Ramachandran plot web server and has several improved options for

displaying the conformation angles in various regions. In addition, options are

provided to display the conformation angles in various secondary structural

elements and regions within the user specified Phi and Psi values in the plot.

The updated version is accessible at the following URL:

http://dicsoft1.physics.iisc.ernet.in/rp/.

PMID: 17897092 [Indexed for MEDLINE]

2816. Protein Pept Lett. 2007;14(7):626-31.

PEPstr: a de novo method for tertiary structure prediction of small bioactive

peptides.

Kaur H(1), Garg A, Raghava GP.

Author information:

(1)Bioinformatics Center, Institute of Microbial Technology, Chandigarh, India.

Among secondary structure elements, beta-turns are ubiquitous and major feature

of bioactive peptides. We analyzed 77 biologically active peptides with length

varying from 9 to 20 residues. Out of 77 peptides, 58 peptides were found to

contain at least one beta-turn. Further, at the residue level, 34.9% of total

peptide residues were found to be in beta-turns, higher than the number of

helical (32.3%) and beta-sheet residues (6.9%). So, we utilized the predicted

beta-turns information to develop an improved method for predicting the

three-dimensional (3D) structure of small peptides. In principle, we built four

different structural models for each peptide. The first 'model I' was built by

assigning all the peptide residues an extended conformation (phi = Psi = 180

degrees ). Second 'model II' was built using the information of regular secondary

structures (helices, beta-strands and coil) predicted from PSIPRED. In third

'model III', secondary structure information including beta-turn types predicted

from BetaTurns method was used. The fourth 'model IV' had main-chain phi, Psi

angles of model III and side chain angles assigned using standard Dunbrack

backbone dependent rotamer library. These models were further refined using AMBER

package and the resultant C(alpha) rmsd values were calculated. It was found that

adding the beta-turns to the regular secondary structures greatly reduces the

rmsd values both before and after the energy minimization. Hence, the results

indicate that regular and irregular secondary structures, particularly beta-turns

information can provide valuable and vital information in the tertiary structure

prediction of small bioactive peptides. Based on the above study, a web server

PEPstr (http://www.imtech.res.in/raghava/pepstr/) was developed for predicting

the tertiary structure of small bioactive peptides.

PMID: 17897087 [Indexed for MEDLINE]

2817. Protein Pept Lett. 2007;14(6):575-80.

Support vector machine based prediction of glutathione S-transferase proteins.

Mishra NK(1), Kumar M, Raghava GP.

Author information:

(1)Bioinformatics Center, Institute of Microbial Technology, Chandigarh, India.

Glutathione S-transferase (GST) proteins play vital role in living organism that

includes detoxification of exogenous and endogenous chemicals, survivability

during stress condition. This paper describes a method developed for predicting

GST proteins. We have used a dataset of 107 GST and 107 non-GST proteins for

training and the performance of the method was evaluated with five-fold

cross-validation technique. First a SVM based method has been developed using

amino acid and dipeptide composition and achieved the maximum accuracy of 91.59%

and 95.79% respectively. In addition we developed a SVM based method using

tripeptide composition and achieved maximum accuracy 97.66% which is better than

accuracy achieved by HMM based searching (96.26%). Based on above study a

web-server GSTPred has been developed

(http://www.imtech.res.in/raghava/gstpred/).

PMID: 17627599 [Indexed for MEDLINE]

2818. BMC Bioinformatics. 2006 Dec 18;7 Suppl 5:S16.

SPIDer: Saccharomyces protein-protein interaction database.

Wu X(1), Zhu L, Guo J, Fu C, Zhou H, Dong D, Li Z, Zhang DY, Lin K.

Author information:

(1)MOE Key Laboratory for Biodiversity Science and Ecological Engineering and

College of Life Sciences, Beijing Normal University, Beijing 100875, China.

wuxm@mail.bnu.edu.cn

BACKGROUND: Since proteins perform their functions by interacting with one

another and with other biomolecules, reconstructing a map of the protein-protein

interactions of a cell, experimentally or computationally, is an important first

step toward understanding cellular function and machinery of a proteome. Solely

derived from the Gene Ontology (GO), we have defined an effective method of

reconstructing a yeast protein interaction network by measuring relative

specificity similarity (RSS) between two GO terms.

DESCRIPTION: Based on the RSS method, here, we introduce a predicted

Saccharomyces protein-protein interaction database called SPIDer. It houses a

gold standard positive dataset (GSP) with high confidence level that covered

79.2% of the high-quality interaction dataset. Our predicted protein-protein

interaction network reconstructed from the GSPs consists of 92,257 interactions

among 3600 proteins, and forms 23 connected components. It also provides general

links to connect predicted protein-protein interactions with three other

databases, DIP, BIND and MIPS. An Internet-based interface provides users with

fast and convenient access to protein-protein interactions based on various

search features (searching by protein information, GO term information or

sequence similarity). In addition, the RSS value of two GO terms in the same

ontology, and the inter-member interactions in a list of proteins of interest or

in a protein complex could be retrieved. Furthermore, the database presents a

user-friendly graphical interface which is created dynamically for visualizing an

interaction sub-network. The database is accessible at

http://cmb.bnu.edu.cn/SPIDer/index.html.

CONCLUSION: SPIDer is a public database server for protein-protein interactions

based on the yeast genome. It provides a variety of search options and graphical

visualization of an interaction network. In particular, it will be very useful

for the study of inter-member interactions among a list of proteins, especially

the protein complex. In addition, based on the predicted interaction dataset,

researchers could analyze the whole interaction network and associate the network

topology with gene/protein properties based on a global or local topology view.

DOI: 10.1186/1471-2105-7-S5-S16

PMCID: PMC1764472

PMID: 17254300 [Indexed for MEDLINE]

2819. BMC Bioinformatics. 2006 Dec 18;7 Suppl 5:S13.

Prediction of the functional class of metal-binding proteins from sequence

derived physicochemical properties by support vector machine approach.

Lin HH(1), Han LY, Zhang HL, Zheng CJ, Xie B, Cao ZW, Chen YZ.

Author information:

(1)Bioinformatics and Drug Design Group, Department of Pharmacy, National

University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543.

honghuang@nus.edu.sg

Metal-binding proteins play important roles in structural stability, signaling,

regulation, transport, immune response, metabolism control, and metal

homeostasis. Because of their functional and sequence diversity, it is desirable

to explore additional methods for predicting metal-binding proteins irrespective

of sequence similarity. This work explores support vector machines (SVM) as such

a method. SVM prediction systems were developed by using 53,333 metal-binding and

147,347 non-metal-binding proteins, and evaluated by an independent set of 31,448

metal-binding and 79,051 non-metal-binding proteins. The computed prediction

accuracy is 86.3%, 81.6%, 83.5%, 94.0%, 81.2%, 85.4%, 77.6%, 90.4%, 90.9%, 74.9%

and 78.1% for calcium-binding, cobalt-binding, copper-binding, iron-binding,

magnesium-binding, manganese-binding, nickel-binding, potassium-binding,

sodium-binding, zinc-binding, and all metal-binding proteins respectively. The

accuracy for the non-member proteins of each class is 88.2%, 99.9%, 98.1%, 91.4%,

87.9%, 94.5%, 99.2%, 99.9%, 99.9%, 98.0%, and 88.0% respectively. Comparable

accuracies were obtained by using a different SVM kernel function. Our method

predicts 67% of the 87 metal-binding proteins non-homologous to any protein in

the Swissprot database and 85.3% of the 333 proteins of known metal-binding

domains as metal-binding. These suggest the usefulness of SVM for facilitating

the prediction of metal-binding proteins. Our software can be accessed at the

SVMProt server http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi.

DOI: 10.1186/1471-2105-7-S5-S13

PMCID: PMC1764469

PMID: 17254297 [Indexed for MEDLINE]

2820. BMC Bioinformatics. 2006 Dec 12;7 Suppl 4:S18.

GenomeBlast: a web tool for small genome comparison.

Lu G(1), Jiang L, Helikar RM, Rowley TW, Zhang L, Chen X, Moriyama EN.

Author information:

(1)Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182, USA.

glu3@mail.unomaha.edu

BACKGROUND: Comparative genomics has become an essential approach for identifying

homologous gene candidates and their functions, and for studying genome

evolution. There are many tools available for genome comparisons. Unfortunately,

most of them are not applicable for the identification of unique genes and the

inference of phylogenetic relationships in a given set of genomes.

RESULTS: GenomeBlast is a Web tool developed for comparative analysis of multiple

small genomes. A new parameter called "coverage" was introduced and used along

with sequence identity to evaluate global similarity between genes. With

GenomeBlast, the following results can be obtained: (1) unique genes in each

genome; (2) homologous gene candidates among compared genomes; (3) 2D plots of

homologous gene candidates along the all pairwise genome comparisons; and (4) a

table of gene presence/absence information and a genome phylogeny. We

demonstrated the functions in GenomeBlast with an example of multiple herpesviral

genome analysis and illustrated how GenomeBlast is useful for small genome

comparison.

CONCLUSION: We developed a Web tool for comparative analysis of small genomes,

which allows the user not only to identify unique genes and homologous gene

candidates among multiple genomes, but also to view their graphical distributions

on genomes, and to reconstruct genome phylogeny. GenomeBlast runs on a Linux

server with 4 CPUs and 4 GB memory. The online version of GenomeBlast is

available to public by using a Web browser with the URL

http://bioinfo-srv1.awh.unomaha.edu/genomeblast/.

DOI: 10.1186/1471-2105-7-S4-S18

PMCID: PMC1780113

PMID: 17217510 [Indexed for MEDLINE]

2821. Autoimmunity. 2006 Dec;39(8):645-50.

PREDNOD, a prediction server for peptide binding to the H-2g7 haplotype of the

non-obese diabetic mouse.

Rajapakse M(1), Zhang GL, Srinivasan KN, Schmidt B, Petrovsky N, Brusic V.

Author information:

(1)Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore,

Singapore, 119613.

The non-obese diabetic (NOD) mouse is a widely used animal model for study of

autoimmune diseases, in particular human type 1 diabetes mellitus (T1DM).

Identification of the subset of peptides that bind MHC molecules comprising the

H-2g7 haplotype of NOD mouse and thereby representing potential NOD T-cell

epitopes is important for research into the pathogenesis and immunotherapy of

T1DM. The H-2g7 haplotype comprises the MHC class-I molecules Kd and Db and a

single class-II molecule I-Ag7. We have developed a prediction system, PREDNOD,

for accurate identification of peptides that bind the MHC molecules constituting

the H-2g7 haplotype. PREDNOD is accessible at

http://antigen.i2r.a-star.edu.sg/Ag7.

DOI: 10.1080/08916930601062494

PMID: 17178561 [Indexed for MEDLINE]

2822. Biochem Biophys Res Commun. 2006 Dec 1;350(4):818-24. Epub 2006 Oct 2.

Prediction of Nepsilon-acetylation on internal lysines implemented in Bayesian

Discriminant Method.

Li A(1), Xue Y, Jin C, Wang M, Yao X.

Author information:

(1)Department of Pathology, School of Medicine, Yale University, New Haven, CT

06520, USA.

Protein acetylation is an important and reversible post-translational

modification (PTM), and it governs a variety of cellular dynamics and plasticity.

Experimental identification of acetylation sites is labor-intensive and often

limited by the availability of reagents such as acetyl-specific antibodies and

optimization of enzymatic reactions. Computational analyses may facilitate the

identification of potential acetylation sites and provide insights into further

experimentation. In this manuscript, we present a novel protein acetylation

prediction program named PAIL, prediction of acetylation on internal lysines,

implemented in a BDM (Bayesian Discriminant Method) algorithm. The accuracies of

PAIL are 85.13%, 87.97%, and 89.21% at low, medium, and high thresholds,

respectively. Both Jack-Knife validation and n-fold cross-validation have been

performed to show that PAIL is accurate and robust. Taken together, we propose

that PAIL is a novel predictor for identification of protein acetylation sites

and may serve as an important tool to study the function of protein acetylation.

PAIL has been implemented in PHP and is freely available on a web server at:

http://bioinformatics.lcd-ustc.org/pail.

DOI: 10.1016/j.bbrc.2006.08.199

PMCID: PMC2093955

PMID: 17045240 [Indexed for MEDLINE]

2823. Bioinformatics. 2006 Dec 1;22(23):2898-904. Epub 2006 Oct 10.

Large scale data mining approach for gene-specific standardization of microarray

gene expression data.

Yoon S(1), Yang Y, Choi J, Seong J.

Author information:

(1)Department of Biological Sciences, Sookmyung Women's University Hyochangwongil

52, Youngsan-gu, Seoul, Republic of Korea, 140-742. yoonsj@sookmyung.ac.kr

MOTIVATION: The identification of the change of gene expression in multifactorial

diseases, such as breast cancer is a major goal of DNA microarray experiments.

Here we present a new data mining strategy to better analyze the marginal

difference in gene expression between microarray samples. The idea is based on

the notion that the consideration of gene's behavior in a wide variety of

experiments can improve the statistical reliability on identifying genes with

moderate changes between samples.

RESULTS: The availability of a large collection of array samples sharing the same

platform in public databases, such as NCBI GEO, enabled us to re-standardize the

expression intensity of a gene using its mean and variation in the wide variety

of experimental conditions. This approach was evaluated via the re-identification

of breast cancer-specific gene expression. It successfully prioritized several

genes associated with breast tumor, for which the expression difference between

normal and breast cancer cells was marginal and thus would have been difficult to

recognize using conventional analysis methods. Maximizing the utility of

microarray data in the public database, it provides a valuable tool particularly

for the identification of previously unrecognized disease-related genes.

AVAILABILITY: A user friendly web-interface

(http://compbio.sookmyung.ac.kr/~lage/) was constructed to provide the present

large-scale approach for the analysis of GEO microarray data (GS-LAGE server).

DOI: 10.1093/bioinformatics/btl500

PMID: 17032674 [Indexed for MEDLINE]

2824. Bioinformatics. 2006 Dec 1;22(23):2948-9. Epub 2006 Oct 4.

FoldUnfold: web server for the prediction of disordered regions in protein chain.

Galzitskaya OV(1), Garbuzynskiy SO, Lobanov MY.

Author information:

(1)Institute of Protein Research, Russian Academy of Sciences 142290, Pushchino,

Moscow Region, Russia. ogalzit@vega.protres.ru

Identification of disordered regions in polypeptide chains is very important

because such regions are essential for protein function. A new parameter, namely

mean packing density of residues has been introduced to detect disordered regions

in a protein sequence. We have demonstrated that regions with weak expected

packing density would be responsible for the appearance of disordered regions.

Our method (FoldUnfold) has been tested on datasets of globular proteins (559

proteins) and long disordered protein segments (129 proteins) and showed improved

performance over some other widely used methods, such as DISOPRED, PONDR VL3H,

IUPred and GlobPlot.AVAILABILITY: The FoldUnfold server is available for users at

http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi. There is a link to our server

through the web site of DisProt (http://www.disprot.org/predictors.php).

DOI: 10.1093/bioinformatics/btl504

PMID: 17021161 [Indexed for MEDLINE]

2825. Bioinformatics. 2006 Dec 1;22(23):2876-82. Epub 2006 Oct 2.

ProtBuD: a database of biological unit structures of protein families and

superfamilies.

Xu Q(1), Canutescu A, Obradovic Z, Dunbrack RL Jr.

Author information:

(1)Institute for Cancer Research, Fox Chase Cancer Center 333 Cottman Avenue,

Philadelphia, PA 19111 USA.

Erratum in

Bioinformatics. 2007 Aug;23(15):2028.

MOTIVATION: Modeling of protein interactions is often possible from known

structures of related complexes. It is often time-consuming to find the most

appropriate template. Hypothesized biological units (BUs) often differ from the

asymmetric units and it is usually preferable to model from the BUs.

RESULTS: ProtBuD is a database of BUs for all structures in the Protein Data Bank

(PDB). We use both the PDBs BUs and those from the Protein Quaternary Server.

ProtBuD is searchable by PDB entry, the Structural Classification of Proteins

(SCOP) designation or pairs of SCOP designations. The database provides the

asymmetric and BU contents of related proteins in the PDB as identified in SCOP

and Position-Specific Iterated BLAST (PSI-BLAST). The asymmetric unit is

different from PDB and/or Protein Quaternary Server (PQS) BUs for 52% of X-ray

structures, and the PDB and PQS BUs disagree on 18% of entries.

AVAILABILITY: The database is provided as a standalone program and a web server

from http://dunbrack.fccc.edu/ProtBuD.php.

DOI: 10.1093/bioinformatics/btl490

PMID: 17018535 [Indexed for MEDLINE]

2826. Cell Immunol. 2006 Dec;244(2):97-100. Epub 2007 Apr 16.

Bioinformatics applied to allergy: allergen databases, from collecting sequence

information to data integration. The Allergome platform as a model.

Mari A(1), Scala E, Palazzo P, Ridolfi S, Zennaro D, Carabella G.

Author information:

(1)Allergy Data Laboratories sc, Via Malipiero 28, 04100 Latina, Italy.

adriano.mari@allergome.org

Allergens are proteins or glycoproteins that are recognized by IgE produced by

the immune system of allergic individuals. Until now around 1,500 allergenic

structures have been identified and this number seems not have reached a plateau

after 3-4 decades of research and the advent of molecular biology. Several

allergen databases are available on Internet. Different aims and philosophies

lead to different products. Here we report about main feature of web sites

dedicated to allergens and we describe in more details our current work on the

Allergome platform. The web server Allergome (www.allergome.org) represent a free

independent open resource whose goal is to provide an exhaustive repository of

data related to all the IgE-binding compounds. The main purpose of Allergome is

to collect a list of allergenic sources and molecules by using the widest

selection criteria and sources. A further development of the Allergome platform

has been represented by the Real Time Monitoring of IgE sensitization module

(ReTiME) that allows uploading of raw data from both in vivo and in vitro

testing, thus representing the first attempt to have IT applied to allergy data

mining. More recently, a new module (RefArray) representing a tool for literature

mining has been released.

DOI: 10.1016/j.cellimm.2007.02.012

PMID: 17434469 [Indexed for MEDLINE]

2827. Genet Mol Res. 2006 Dec 1;5(4):717-22.

The Star STING server: a multiplatform environment for protein structure

analysis.

Neshich G(1), Mazoni I, Oliveira SR, Yamagishi ME, Kuser-Falcão PR, Borro LC,

Morita DU, Souza KR, Almeida GV, Rodrigues DN, Jardine JG, Togawa RC, Mancini AL,

Higa RH, Cruz SA, Vieira FD, Santos EH, Melo RC, Santoro MM.

Author information:

(1)Núcleo de Bioinformática Estrutural, Embrapa Informática Agropecuária,

Campinas, SP, Brasil. neshich@cbi.cnptia.embrapa.br

Star STING is the latest version of the STING suite of programs and corresponding

database. We report on five important aspects of this package that have acquired

some new characteristics, designed to add key advantages to the whole suite: 1)

availability for most popular platforms and browsers, 2) introduction of the

STING\_DB quality assessment, 3) improvement in algorithms for calculation of

three STING parameters, 4) introduction of five new STING modules, and 5)

expansion of the existing modules. Star STING is freely accessible at:

http://sms.cbi.cnptia.embrapa.br/SMS/, http://trantor.bioc.columbia.edu/SMS,

http://www.es.embnet.org/SMS/, http://gibk26.bse.kyutech.ac.jp/SMS/ and

http://www.ar.embnet.org/SMS.

PMID: 17183482 [Indexed for MEDLINE]

2828. J Proteome Res. 2006 Dec;5(12):3420-8.

Large-scale predictions of gram-negative bacterial protein subcellular locations.

Chou KC(1), Shen HB.

Author information:

(1)Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego,

California 92130, USA. kchou@san.rr.com

Many species of Gram-negative bacteria are pathogenic bacteria that can cause

disease in a host organism. This pathogenic capability is usually associated with

certain components in Gram-negative cells. Therefore, developing an automated

method for fast and reliable prediction of Gram-negative protein subcellular

location will allow us to not only timely annotate gene products, but also screen

candidates for drug discovery. However, protein subcellular location prediction

is a very difficult problem, particularly when more location sites need to be

involved and when unknown query proteins do not have significant homology to

proteins of known subcellular locations. PSORT-B, a recently updated version of

PSORT, widely used for predicting Gram-negative protein subcellular location,

only covers five location sites. Also, the data set used to train PSORT-B

contains many proteins with high degrees of sequence identity in a same location

group and, hence, may bear a strong homology bias. To overcome these problems, a

new predictor, called "Gneg-PLoc", is developed. Featured by fusing many basic

classifiers each being trained with a stringent data set containing proteins with

strictly less than 25% sequence identity to one another in a same location group,

the new predictor can cover eight subcellular locations; that is, cytoplasm,

extracellular space, fimbrium, flagellum, inner membrane, nucleoid, outer

membrane, and periplasm. In comparison with PSORT-B, the new predictor not only

covers more subcellular locations, but also yields remarkably higher success

rates. Gneg-PLoc is available as a Web server at

http://202.120.37.186/bioinf/Gneg. To support the demand of people working in the

relevant areas, a downloadable file is provided at the same Web site to list the

results identified by Gneg-PLoc for 49 907 Gram-negative protein entries in the

Swiss-Prot database that have no subcellular location annotations or are

annotated with uncertain terms. The large-scale results will be updated twice a

year to cover the new entries of Gram-negative bacterial proteins and reflect the

new development of Gneg-PLoc.

DOI: 10.1021/pr060404b

PMID: 17137343 [Indexed for MEDLINE]

2829. Neural Comput. 2006 Dec;18(12):2923-7.

A distributed computing tool for generating neural simulation databases.

Calin-Jageman RJ(1), Katz PS.

Author information:

(1)rcalinjageman@gsu.edu

After developing a model neuron or network, it is important to systematically

explore its behavior across a wide range of parameter values or experimental

conditions, or both. However, compiling a very large set of simulation runs is

challenging because it typically requires both access to and expertise with

high-performance computing facilities. To lower the barrier for large-scale model

analysis, we have developed NeuronPM, a client/server application that creates a

"screen-saver" cluster for running simulations in NEURON (Hines & Carnevale,

1997). NeuronPM provides a user-friendly way to use existing computing resources

to catalog the performance of a neural simulation across a wide range of

parameter values and experimental conditions. The NeuronPM client is a

Windows-based screen saver, and the NeuronPM server can be hosted on any

Apache/PHP/MySQL server. During idle time, the client retrieves model files and

work assignments from the server, invokes NEURON to run the simulation, and

returns results to the server. Administrative panels make it simple to upload

model files, define the parameters and conditions to vary, and then monitor

client status and work progress. NeuronPM is open-source freeware and is

available for download at http://neuronpm.homeip.net . It is a useful entry-level

tool for systematically analyzing complex neuron and network simulations.

DOI: 10.1162/neco.2006.18.12.2923

PMID: 17052151 [Indexed for MEDLINE]

2830. PLoS Comput Biol. 2006 Nov 17;2(11):e155. Epub 2006 Oct 5.

3D complex: a structural classification of protein complexes.

Levy ED(1), Pereira-Leal JB, Chothia C, Teichmann SA.

Author information:

(1)Medical Research Council Laboratory of Molecular Biology, Cambridge, United

Kingdom. elevy@mrc-lmb.cam.ac.uk

Most of the proteins in a cell assemble into complexes to carry out their

function. It is therefore crucial to understand the physicochemical properties as

well as the evolution of interactions between proteins. The Protein Data Bank

represents an important source of information for such studies, because more than

half of the structures are homo- or heteromeric protein complexes. Here we

propose the first hierarchical classification of whole protein complexes of known

3-D structure, based on representing their fundamental structural features as a

graph. This classification provides the first overview of all the complexes in

the Protein Data Bank and allows nonredundant sets to be derived at different

levels of detail. This reveals that between one-half and two-thirds of known

structures are multimeric, depending on the level of redundancy accepted. We also

analyse the structures in terms of the topological arrangement of their subunits

and find that they form a small number of arrangements compared with all

theoretically possible ones. This is because most complexes contain four subunits

or less, and the large majority are homomeric. In addition, there is a strong

tendency for symmetry in complexes, even for heteromeric complexes. Finally,

through comparison of Biological Units in the Protein Data Bank with the Protein

Quaternary Structure database, we identified many possible errors in quaternary

structure assignments. Our classification, available as a database and Web server

at http://www.3Dcomplex.org, will be a starting point for future work aimed at

understanding the structure and evolution of protein complexes.

DOI: 10.1371/journal.pcbi.0020155

PMCID: PMC1636673

PMID: 17112313 [Indexed for MEDLINE]

2831. BMC Bioinformatics. 2006 Nov 16;7:503.

Predicting residue contacts using pragmatic correlated mutations method: reducing

the false positives.

Kundrotas PJ(1), Alexov EG.

Author information:

(1)Computational Biophysics and Bioinformatics, Department of Physics, Clemson

University, Clemson, SC 29634, USA. pkundro@clemson.edu <pkundro@clemson.edu>

BACKGROUND: Predicting residues' contacts using primary amino acid sequence alone

is an important task that can guide 3D structure modeling and can verify the

quality of the predicted 3D structures. The correlated mutations (CM) method

serves as the most promising approach and it has been used to predict amino acids

pairs that are distant in the primary sequence but form contacts in the native 3D

structure of homologous proteins.

RESULTS: Here we report a new implementation of the CM method with an added set

of selection rules (filters). The parameters of the algorithm were optimized

against fifteen high resolution crystal structures with optimization criterion

that maximized the confidentiality of the predictions. The optimization resulted

in a true positive ratio (TPR) of 0.08 for the CM without filters and a TPR of

0.14 for the CM with filters. The protocol was further benchmarked against 65

high resolution structures that were not included in the optimization test. The

benchmarking resulted in a TPR of 0.07 for the CM without filters and to a TPR of

0.09 for the CM with filters.

CONCLUSION: Thus, the inclusion of selection rules resulted to an overall

improvement of 30%. In addition, the pair-wise comparison of TPR for each protein

without and with filters resulted in an average improvement of 1.7. The

methodology was implemented into a web server

http://www.ces.clemson.edu/compbio/recon that is freely available to the public.

The purpose of this implementation is to provide the 3D structure predictors with

a tool that can help with ranking alternative models by satisfying the largest

number of predicted contacts, as well as it can provide a confidence score for

contacts in cases where structure is known.

DOI: 10.1186/1471-2105-7-503

PMCID: PMC1654194

PMID: 17109752 [Indexed for MEDLINE]

2832. Bioinformatics. 2006 Nov 15;22(22):2838-40. Epub 2006 Oct 10.

Integration of gel-based proteome data with pProRep.

Laukens K(1), Matthiesen R, Lemière F, Esmans E, Onckelen HV, Jensen ON, Witters

E.

Author information:

(1)Laboratory of Plant Physiology and Plant Biochemistry, University of Antwerp

Groenenborgerlaan 171, B-2020 Antwerp, Belgium. kris.laukens@ua.ac.be

pProRep is a web application integrating electrophoretic and mass spectral data

from proteome analyses into a relational database. The graphical web-interface

allows users to upload, analyse and share experimental proteome data. It offers

researchers the possibility to query all previously analysed datasets and can

visualize selected features, such as the presence of a certain set of ions in a

peptide mass spectrum, on the level of the two-dimensional gel.AVAILABILITY: The

pProRep package and instructions for its use can be downloaded from

http://www.ptools.ua.ac.be/pProRep. The application requires a web server that

runs PHP 5 (http://www.php.net) and MySQL. Some (non-essential) extensions need

additional freely available libraries: details are described in the installation

instructions.

DOI: 10.1093/bioinformatics/btl487

PMID: 17032679 [Indexed for MEDLINE]

2833. BMC Bioinformatics. 2006 Nov 3;7:485.

Machine learning techniques in disease forecasting: a case study on rice blast

prediction.

Kaundal R(1), Kapoor AS, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A,

Chandigarh 160036, India. rakesh@imtech.res.in <rakesh@imtech.res.in>

BACKGROUND: Diverse modeling approaches viz. neural networks and multiple

regression have been followed to date for disease prediction in plant

populations. However, due to their inability to predict value of unknown data

points and longer training times, there is need for exploiting new prediction

softwares for better understanding of plant-pathogen-environment relationships.

Further, there is no online tool available which can help the plant researchers

or farmers in timely application of control measures. This paper introduces a new

prediction approach based on support vector machines for developing weather-based

prediction models of plant diseases.

RESULTS: Six significant weather variables were selected as predictor variables.

Two series of models (cross-location and cross-year) were developed and validated

using a five-fold cross validation procedure. For cross-year models, the

conventional multiple regression (REG) approach achieved an average correlation

coefficient (r) of 0.50, which increased to 0.60 and percent mean absolute error

(%MAE) decreased from 65.42 to 52.24 when back-propagation neural network (BPNN)

was used. With generalized regression neural network (GRNN), the r increased to

0.70 and %MAE also improved to 46.30, which further increased to r = 0.77 and

%MAE = 36.66 when support vector machine (SVM) based method was used. Similarly,

cross-location validation achieved r = 0.48, 0.56 and 0.66 using REG, BPNN and

GRNN respectively, with their corresponding %MAE as 77.54, 66.11 and 58.26. The

SVM-based method outperformed all the three approaches by further increasing r to

0.74 with improvement in %MAE to 44.12. Overall, this SVM-based prediction

approach will open new vistas in the area of forecasting plant diseases of

various crops.

CONCLUSION: Our case study demonstrated that SVM is better than existing machine

learning techniques and conventional REG approaches in forecasting plant

diseases. In this direction, we have also developed a SVM-based web server for

rice blast prediction, a first of its kind worldwide, which can help the plant

science community and farmers in their decision making process. The server is

freely available at http://www.imtech.res.in/raghava/rbpred/.

DOI: 10.1186/1471-2105-7-485

PMCID: PMC1647291

PMID: 17083731 [Indexed for MEDLINE]

2834. Bioinformatics. 2006 Nov 1;22(21):2693-4. Epub 2006 Aug 29.

iFold: a platform for interactive folding simulations of proteins.

Sharma S(1), Ding F, Nie H, Watson D, Unnithan A, Lopp J, Pozefsky D, Dokholyan

NV.

Author information:

(1)Department of Biochemistry and Biophysics, University of North Carolina Chapel

Hill, NC 27599, USA.

We built a novel web-based platform for performing discrete molecular dynamics

simulations of proteins. In silico protein folding involves searching for minimal

frustration in the vast conformational landscape. Conventional approaches for

simulating protein folding insufficiently address the problem of simulations in

relevant time and length scales necessary for a mechanistic understanding of

underlying biomolecular phenomena. Discrete molecular dynamics (DMD) offers an

opportunity to bridge the size and timescale gaps and uncover the structural and

biological properties of experimentally undetectable protein dynamics. The iFold

server supports large-scale simulations of protein folding, thermal denaturation,

thermodynamic scan, simulated annealing and p(fold) analysis using DMD and

coarse-grained protein model with structure-based Gō-interactions between amino

acids.AVAILABILITY: http://ifold.dokhlab.org

DOI: 10.1093/bioinformatics/btl460

PMID: 16940324 [Indexed for MEDLINE]

2835. Bioinformatics. 2006 Nov 1;22(21):2691-2. Epub 2006 Aug 23.

MMM: a sequence-to-structure alignment protocol.

Rai BK(1), Madrid-Aliste CJ, Fajardo JE, Fiser A.

Author information:

(1)Department of Biochemistry and Seaver Center for Bioinformatics, Albert

Einstein College of Medicine 1300 Morris Park Avenue, Bronx, NY 10461, USA.

MOTIVATION: Accurate alignment of a target sequence to a template structure

continues to be a bottleneck in producing good quality comparative protein

structure models.

RESULTS: Multiple Mapping Method (MMM) is a comparative protein structure

modeling server with an emphasis on a novel alignment optimization protocol. MMM

takes inputs from five profile-to-profile based alignment methods. The

alternatively aligned regions from the input alignment set are combined according

to their fit in the structural environment of the template structure. The

resulting, optimally spliced MMM alignment is used as input to an automated

comparative modeling module to produce a full atom model.

AVAILABILITY: The MMM server is freely accessible at

http://www.fiserlab.org/servers/mmm

DOI: 10.1093/bioinformatics/btl449

PMID: 16928737 [Indexed for MEDLINE]

2836. Bioinformatics. 2006 Nov 1;22(21):2619-27. Epub 2006 Aug 23.

Anisotropic network model: systematic evaluation and a new web interface.

Eyal E(1), Yang LW, Bahar I.

Author information:

(1)Department of Computational Biology, School of Medicine, University of

Pittsburgh, PA 15213, USA.

MOTIVATION: The Anisotropic Network Model (ANM) is a simple yet powerful model

for normal mode analysis of proteins. Despite its broad use for exploring

biomolecular collective motions, ANM has not been systematically evaluated to

date. A lack of a convenient interface has been an additional obstacle for easy

usage.

RESULTS: ANM has been evaluated on a large set of proteins to establish the

optimal model parameters that achieve the highest correlation with experimental

data and its limits of accuracy and applicability. Residue fluctuations in

globular proteins are shown to be more accurately predicted than those in

nonglobular proteins, and core residues are more accurately described than

solvent-exposed ones. Significant improvement in agreement with experiments is

observed with increase in the resolution of the examined structure. A new server

for ANM calculations is presented, which offers flexible options for controlling

model parameters and output formats, interactive animation of collective modes

and advanced graphical features.

AVAILABILITY: ANM server (http://www.ccbb.pitt.edu/anm)

DOI: 10.1093/bioinformatics/btl448

PMID: 16928735 [Indexed for MEDLINE]

2837. Bioinformatics. 2006 Nov 1;22(21):2702-3. Epub 2006 Aug 23.

TESD: a transposable element dynamics simulation environment.

Deceliere G(1), Letrillard Y, Charles S, Biémont C.

Author information:

(1)Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS, Université

Claude Bernard Lyon 1 69622, Villeurbanne Cedex, France.

Various mathematical models have been used to explore the dynamics of

transposable elements (TEs) within their host genomes. However, numerous factors

can influence their dynamics, and we know only little about the dynamics of TEs

when they first began to invade populations. In addition, the influence of

population structuring has only recently been investigated. Transposable Element

Simulator Dynamics, a population genomics simulation environment, has therefore

been developed to provide a simple tool for analyzing the dynamics of TEs in a

community based on (i) various TE parameters, such as the transposition and

excision rates, the recombination rate and the coefficient of selection against

TE insertions; and (ii) population parameters, such as population size and

migration rates. The simulations can be used to illustrate the dynamic fate of

TEs in structured populations, can be extended by using more specific molecular

or demographic models, and can be useful for teaching population genetics and

genomics.AVAILABILITY: TESD is distributed under GPL from the Pôle

Bioinformatique Lyonnais (PBIL) web server at

http://pbil.univ-lyon1.fr/software/TESD

DOI: 10.1093/bioinformatics/btl454

PMID: 16928734 [Indexed for MEDLINE]

2838. Eur Spine J. 2006 Nov;15(11):1687-94. Epub 2006 May 20.

Internet based multicenter study for thoracolumbar injuries: a new concept and

preliminary results.

Knop C(1), Reinhold M, Roeder C, Staub L, Schmid R, Beisse R, Bühren V, Blauth M.

Author information:

(1)Trauma Surgery and Sports Medicine, Innsbruck Medical University, Innsbruck,

Austria. christian.knop@uibk.ac.at

This article reports about the internet based, second multicenter study (MCS II)

of the spine study group (AG WS) of the German trauma association (DGU). It

represents a continuation of the first study conducted between the years 1994 and

1996 (MCS I). For the purpose of one common, centralised data capture

methodology, a newly developed internet-based data collection system (

http://www.memdoc.org ) of the Institute for Evaluative Research in Orthopaedic

Surgery of the University of Bern was used. The aim of this first publication on

the MCS II was to describe in detail the new method of data collection and the

structure of the developed data base system, via internet. The goal of the study

was the assessment of the current state of treatment for fresh traumatic injuries

of the thoracolumbar spine in the German speaking part of Europe. For that

reason, we intended to collect large number of cases and representative, valid

information about the radiographic, clinical and subjective treatment outcomes.

Thanks to the new study design of MCS II, not only the common surgical treatment

concepts, but also the new and constantly broadening spectrum of spine surgery,

i.e. vertebro-/kyphoplasty, computer assisted surgery and navigation,

minimal-invasive, and endoscopic techniques, documented and evaluated. We present

a first statistical overview and preliminary analysis of 18 centers from Germany

and Austria that participated in MCS II. A real time data capture at source was

made possible by the constant availability of the data collection system via

internet access. Following the principle of an application service provider,

software, questionnaires and validation routines are located on a central server,

which is accessed from the periphery (hospitals) by means of standard Internet

browsers. By that, costly and time consuming software installation and

maintenance of local data repositories are avoided and, more importantly,

cumbersome migration of data into one integrated database becomes obsolete.

Finally, this set-up also replaces traditional systems wherein paper

questionnaires were mailed to the central study office and entered by hand

whereby incomplete or incorrect forms always represent a resource consuming

problem and source of error. With the new study concept and the expanded

inclusion criteria of MCS II 1, 251 case histories with admission and surgical

data were collected. This remarkable number of interventions documented during 24

months represents an increase of 183% compared to the previously conducted MCS I.

The concept and technical feasibility of the MEMdoc data collection system was

proven, as the participants of the MCS II succeeded in collecting data ever

published on the largest series of patients with spinal injuries treated within a

2 year period.

DOI: 10.1007/s00586-006-0135-7

PMID: 16715307 [Indexed for MEDLINE]

2839. Pathologe. 2006 Nov;27(6):469-76.

[Virtual microscopy: first applications].

[Article in German]

Glatz-Krieger K(1), Glatz D, Mihatsch MJ.

Author information:

(1)Institut für Pathologie, Universitätsspital Basel, Schweiz.

Katharina.Glatz@unibas.ch

Only recently fast-paced developments in computer technology allowed for the

digitization of complete histologic slides. The resulting virtual slides may be

viewed via webbrowser by any number of pathologists or students independent of

time and location. Usage of a virtual microscope simply requires a computer

workstation with a fast internet connection, which opens this technology to a

broad public. A virtual microscopy system consists of three components:

acquisition, server and client. Such systems are under development by different

commercial and academic bodies worldwide. We have developed a virtual microscope

system called vMic (http://www.vmic.unibas.ch) which provides virtual slides of

very high image quality. Several successfully held online slide seminars and a

histology course for students in dentistry are freely accessible in the internet.

With the commercial availability of ultra rapid and easy-to-use slide scanners

and the fast improvements of technology virtual microscopy will offer many

applications in teaching, research and diagnostics. Thanks to additional

functionalities, real microscopes will most likely be replaced by computer

workstations in a couple of years.

DOI: 10.1007/s00292-005-0782-1

PMID: 16096757 [Indexed for MEDLINE]

2840. Zhongguo Yi Liao Qi Xie Za Zhi. 2006 Nov;30(6):419-21, 456.

[A web-based biomedical image mosaicing system].

[Article in Chinese]

Zhang M(1), Yan ZZ, Pan ZJ, Shao SJ.

Author information:

(1)School of Communication and Information Engineering, Shanghai University.

This paper describes a web service for biomedical image mosaicing. A web site

based on CGI (Common Gateway Interface) is implemented. The system is based on

Browser/Server model and is tested in www. Finally implementation examples and

experiment results are provided.

PMID: 17300008 [Indexed for MEDLINE]

2841. J Theor Biol. 2006 Oct 21;242(4):941-6. Epub 2006 May 16.

Using stacked generalization to predict membrane protein types based on

pseudo-amino acid composition.

Wang SQ(1), Yang J, Chou KC.

Author information:

(1)Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong

University, Shanghai 200240, China. lifesci-sjtu@san.rr.com

<lifesci-sjtu@san.rr.com>

Membrane proteins are vitally important for many biological processes and have

become an attractive target for both basic research and drug design. Knowledge of

membrane protein types often provides useful clues in deducing the functions of

uncharacterized membrane proteins. With the unprecedented increasing of newly

found protein sequences in the post-genomic era, it is highly demanded to develop

an automated method for fast and accurately identifying the types of membrane

proteins according to their amino acid sequences. Although quite a few

identifiers have been developed in this regard through various approaches, such

as covariant discriminant (CD), support vector machine (SVM), artificial neural

network (ANN), and K-nearest neighbor (KNN), classifier the way they operate the

identification is basically individual. As is well known, wise persons usually

take into account the opinions from several experts rather than rely on only one

when they are making critical decisions. Likewise, a sophisticated identifier

should be trained by several different modes. In view of this, based on the frame

of pseudo-amino acid that can incorporate a considerable amount of sequence-order

effects, a novel approach called "stacked generalization" or "stacking" has been

introduced. Unlike the "bagging" and "boosting" approaches which only combine the

classifiers of a same type, the stacking approach can combine several different

types of classifiers through a meta-classifier to maximize the generalization

accuracy. The results thus obtained were very encouraging. It is anticipated that

the stacking approach may also hold a high potential to improve the

identification quality for, among many other protein attributes, subcellular

location, enzyme family class, protease type, and protein-protein interaction

type. The stacked generalization classifier is available as a web-server named

"SG-MPt\_Pred" at: http://202.120.37.186/bioinf/wangsq/service.htm.

DOI: 10.1016/j.jtbi.2006.05.006

PMID: 16806277 [Indexed for MEDLINE]

2842. BMC Bioinformatics. 2006 Oct 5;7:431.

The BiSearch web server.

Arányi T(1), Váradi A, Simon I, Tusnády GE.

Author information:

(1)lnstitute of Enzymology, BRC, HAS, H-1113 Karolina út 29, Budapest, Hungary.

aranyi@enzim.hu

BACKGROUND: A large number of PCR primer-design softwares are available online.

However, only very few of them can be used for the design of primers to amplify

bisulfite-treated DNA templates, necessary to determine genomic DNA methylation

profiles. Indeed, the number of studies on bisulfite-treated templates

exponentially increases as determining DNA methylation becomes more important in

the diagnosis of cancers. Bisulfite-treated DNA is difficult to amplify since

undesired PCR products are often amplified due to the increased sequence

redundancy after the chemical conversion. In order to increase the efficiency of

PCR primer-design, we have developed BiSearch web server, an online primer-design

tool for both bisulfite-treated and native DNA templates.

RESULTS: The web tool is composed of a primer-design and an electronic PCR (ePCR)

algorithm. The completely reformulated ePCR module detects potential mispriming

sites as well as undesired PCR products on both cDNA and native or

bisulfite-treated genomic DNA libraries. Due to the new algorithm of the current

version, the ePCR module became approximately hundred times faster than the

previous one and gave the best performance when compared to other web based

tools. This high-speed ePCR analysis made possible the development of the new

option of high-throughput primer screening. BiSearch web server can be used for

academic researchers at the http://bisearch.enzim.hu site.

CONCLUSION: BiSearch web server is a useful tool for primer-design for any DNA

template and especially for bisulfite-treated genomes. The ePCR tool for fast

detection of mispriming sites and alternative PCR products in cDNA libraries and

native or bisulfite-treated genomes are the unique features of the new version of

BiSearch software.

DOI: 10.1186/1471-2105-7-431

PMCID: PMC1609187

PMID: 17022803 [Indexed for MEDLINE]

2843. Bioinformatics. 2006 Oct 1;22(19):2439-40.

APDB: a web server to evaluate the accuracy of sequence alignments using

structural information.

Armougom F(1), Poirot O, Moretti S, Higgins DG, Bucher P, Keduas V, Notredame C.

Author information:

(1)CNRS UPR2589, Institute for Structural Biology and Microbiology (IBSM), Parc

Scientifique de Luminy, 163 Avenue de Luminy, FR-13288, Marseille cedex 09,

France.

The APDB webserver uses structural information to evaluate the alignment of

sequences with known structures. It returns a score correlated to the overall

alignment accuracy as well as a local evaluation. Any sequence alignment can be

analyzed with APDB provided it includes at least two proteins with known

structures. Sequences without a known structure are simply ignored and do not

contribute to the scoring procedure.AVAILABILITY: APDB is part of the T-Coffee

suite of tools for alignment analysis, it is available on www.tcoffee.org. A

stand-alone version of the package is also available as a freeware open source

from the same address.

DOI: 10.1093/bioinformatics/btl404

PMID: 17032685 [Indexed for MEDLINE]

2844. Bioinformatics. 2006 Oct 1;22(19):2441-3. Epub 2006 Jul 26.

JVirGel 2.0: computational prediction of proteomes separated via two-dimensional

gel electrophoresis under consideration of membrane and secreted proteins.

Hiller K(1), Grote A, Maneck M, Münch R, Jahn D.

Author information:

(1)Institut für Mikrobiologie, Technische Universität, Braunschweig

Spielmannstrasse 7, D-38106 Braunschweig, Germany.

MOTIVATION: After the publication of JVirGel 1.0 in 2003 we got many requests and

suggestions from the proteomics community to further improve the performance of

the software and to add additional useful new features.

RESULTS: The integration of the PrediSi algorithm for the prediction of signal

peptides for the Sec-dependent protein export into JVirGel 2.0 allows the

exclusion of most exported preproteins from calculated proteomic maps and

provides the basis for the calculation of Sec-based secretomes. A tool for the

identification of transmembrane helices carrying proteins (JCaMelix) and the

prediction of the corresponding membrane proteome was added. Finally, in order to

directly compare experimental and calculated proteome data, a function to overlay

and evaluate predicted and experimental two-dimensional gels was included.

AVAILABILITY: JVirGel 2.0 is freely available as precompiled package for the

installation on Windows or Linux operating systems. Furthermore, there is a

completely platform-independent Java version available for download.

Additionally, we provide a Java Server Pages based version of JVirGel 2.0 which

can be operated in nearly all web browsers. All versions are accessible at

http://www.jvirgel.de

DOI: 10.1093/bioinformatics/btl409

PMID: 16870933 [Indexed for MEDLINE]

2845. J Proteome Res. 2006 Oct;5(10):2849-52.

GAPP: a fully automated software for the confident identification of human

peptides from tandem mass spectra.

Shadforth I(1), Xu W, Crowther D, Bessant C.

Author information:

(1)Cranfield University, Silsoe, Bedfordshire MK45 4DT, United Kingdom.

This paper introduces the genome annotating proteomic pipeline (GAPP), a totally

automated publicly available software pipeline for the identification of peptides

and proteins from human proteomic tandem mass spectrometry data. The pipeline

takes as its input a series of MS/MS peak lists from a given experimental sample

and produces a series of database entries corresponding to the peptides observed

within the sample, along with related confidence scores. The pipeline is capable

of finding any peptides expected, including those that cross intron-exon

boundaries, and those due to single nucleotide polymorphisms (SNPs), alternate

splicing, and post-translational modifications (PTMs). GAPP can therefore be used

to re-annotate genomes, and this is supported through the inclusion of a

Distributed Annotation System (DAS) server, which allows the peptides identified

by the pipeline to be displayed in their genomic context within the Ensembl

genome browser. GAPP is freely available via the web, at www. gapp.info.

DOI: 10.1021/pr060205s

PMID: 17022656 [Indexed for MEDLINE]

2846. J Struct Biol. 2006 Oct;156(1):230-43. Epub 2006 May 7.

Modeling AAA+ ring complexes from monomeric structures.

Diemand AV(1), Lupas AN.

Author information:

(1)Department of Protein Evolution, Max-Planck-Institute for Developmental

Biology, D-72076 Tübingen, Germany.

AAA+ proteins form large, ring-shaped complexes, which act as energy-dependent

unfoldases of macromolecules. Many crystal structures of proteins in this

superfamily have been determined, but mostly in monomeric or non-physiological

oligomeric forms. The assembly of ring-shaped complexes from monomer coordinates

is, therefore, of considerable interest. We have extracted structural features of

complex formation relating to the distance of monomers from the central axis,

their relative orientation and the molecular contacts at their interfaces from

experimentally determined oligomers and have implemented a semi-automated

modeling procedure based on RosettaDock into the iMolTalk server

(http://protevo.eb.tuebingen.mpg.de/iMolTalk). As examples of this procedure, we

present here models of Apaf-1, MalT and ClpB. We show that the recent EM-based

model of the apoptosome is not compatible with the conserved structural features

of AAA+ complexes and that the D1 and D2 rings of ClpB are most likely offset by

one subunit, in agreement with the structure proposed for ClpA.

DOI: 10.1016/j.jsb.2006.04.011

PMID: 16765605 [Indexed for MEDLINE]

2847. Proteins. 2006 Oct 1;65(1):49-54.

Better prediction of the location of alpha-turns in proteins with support vector

machine.

Wang Y(1), Xue Z, Xu J.

Author information:

(1)Department of Control Science and Engineering, Huazhong University of Science

and Technology, Wuhan City, China.

We have developed a novel method named AlphaTurn to predict alpha-turns in

proteins based on the support vector machine (SVM). The prediction was done on a

data set of 469 nonhomologous proteins containing 967 alpha-turns. A great

improvement in prediction performance was achieved by using multiple sequence

alignment generated by PSI-BLAST as input instead of the single amino acid

sequence. The introduction of secondary structure information predicted by

PSIPRED also improved the prediction performance. Moreover, we handled the very

uneven data set by combining the cost factor j with the "state-shifting" rule.

This further promoted the prediction quality of our method. The final SVM model

yielded a Matthews correlation coefficient (MCC) of 0.25 by a 10-fold

cross-validation. To our knowledge, this MCC value is the highest obtained so far

for predicting alpha-turns. An online Web server based on this method has been

developed and can be freely accessed at http://bmc.hust.edu.cn/bioinformatics/ or

http://210.42.106.80/.

Proteins 2006. (c) 2006 Wiley-Liss, Inc.

DOI: 10.1002/prot.21062

PMID: 16894602 [Indexed for MEDLINE]

2848. Proteomics. 2006 Oct;6(19):5099-105.

GNBSL: a new integrative system to predict the subcellular location for

Gram-negative bacteria proteins.

Guo J(1), Lin Y, Liu X.

Author information:

(1)Department of Mathematical Sciences, Laboratory of Statistical Computing &

Bioinformatics, Tsinghua University, Beijing, P R China. genovo@126.com

This paper proposes a new integrative system (GNBSL--Gram-negative bacteria

subcellular localization) for subcellular localization specifized on the

Gram-negative bacteria proteins. First, the system generates a position-specific

frequency matrix (PSFM) and a position-specific scoring matrix (PSSM) for each

protein sequence by searching the Swiss-Prot database. Then different features

are extracted by four modules from the PSFM and the PSSM. The features include

whole-sequence amino acid composition, N- and C-terminus amino acid composition,

dipeptide composition, and segment composition. Four probabilistic neural network

(PNN) classifiers are used to classify these modules. To further improve the

performance, two modules trained by support vector machine (SVM) are added in

this system. One module extracts the residue-couple distribution from the amino

acid sequence and the other module applies a pairwise profile alignment kernel to

measure the local similarity between every two sequences. Finally, an additional

SVM is used to fuse the outputs from the six modules. Test on a benchmark dataset

shows that the overall success rate of GNBSL is higher than those of PSORT-B,

CELLO, and PSLpred. A web server GNBSL can be visited from

http://166.111.24.5/webtools/GNBSL/index.htm.

DOI: 10.1002/pmic.200600064

PMID: 16955516 [Indexed for MEDLINE]

2849. Bioinformatics. 2006 Sep 15;22(18):2291-7. Epub 2006 Jul 14.

Global topological features of cancer proteins in the human interactome.

Jonsson PF(1), Bates PA.

Author information:

(1)Biomolecular Modelling Laboratory, Cancer Research UK London Research

Institute 44 Lincoln's Inn Fields, London, WC2A 3PX, UK.

MOTIVATION: The study of interactomes, or networks of protein-protein

interactions, is increasingly providing valuable information on biological

systems. Here we report a study of cancer proteins in an extensive human

protein-protein interaction network constructed by computational methods.

RESULTS: We show that human proteins translated from known cancer genes exhibit a

network topology that is different from that of proteins not documented as being

mutated in cancer. In particular, cancer proteins show an increase in the number

of proteins they interact with. They also appear to participate in central hubs

rather than peripheral ones, mirroring their greater centrality and participation

in networks that form the backbone of the proteome. Moreover, we show that cancer

proteins contain a high ratio of highly promiscuous structural domains, i.e.,

domains with a high propensity for mediating protein interactions. These

observations indicate an underlying evolutionary distinction between the two

groups of proteins, reflecting the central roles of proteins, whose mutations

lead to cancer.

CONTACT: paul.bates@cancer.org.uk

SUPPLEMENTARY INFORMATION: The interactome data are available though the PIP

(Potential Interactions of Proteins) web server at

http://bmm.cancerresearchuk.org/servers/pip. Further additional material is

available at http://bmm.cancerresearchuk.org/servers/pip/bioinformatics/

DOI: 10.1093/bioinformatics/btl390

PMCID: PMC1865486

PMID: 16844706 [Indexed for MEDLINE]

2850. Bioinformatics. 2006 Sep 15;22(18):2310-2. Epub 2006 Apr 13.

Dragon Promoter Mapper (DPM): a Bayesian framework for modelling promoter

structures.

Chowdhary R(1), Tan SL, Ali RA, Boerlage B, Wong L, Bajic VB.

Author information:

(1)Knowledge Extraction Lab, Institute for Infocomm Research 21 Heng Mui Keng

Terrace, Singapore 119613, Singapore.

Dragon Promoter Mapper (DPM) is a tool to model promoter structure of

co-regulated genes using methodology of Bayesian networks. DPM exploits an

exhaustive set of motif features (such as motif, its strand, the order of motif

occurrence and mutual distance between the adjacent motifs) and generates models

from the target promoter sequences, which may be used to (1) detect regions in a

genomic sequence which are similar to the target promoters or (2) to classify

other promoters as similar or not to the target promoter group. DPM can also be

used for modelling of enhancers and silencers.AVAILABILITY:

http://defiant.i2r.a-star.edu.sg/projects/BayesPromoter/

CONTACT: vlad@sanbi.ac.za

SUPPLEMENTARY INFORMATION: Manual for using DPM web server is provided at

http://defiant.i2r.a-star.edu.sg/projects/BayesPromoter/html/manual/manual.htm.

DOI: 10.1093/bioinformatics/btl125

PMID: 16613910 [Indexed for MEDLINE]

2851. BMC Bioinformatics. 2006 Sep 6;7:404.

RDMAS: a web server for RNA deleterious mutation analysis.

Shu W(1), Bo X, Liu R, Zhao D, Zheng Z, Wang S.

Author information:

(1)Beijing Institute of Radiation Medicine, Beijing 100850, China.

shuwj@bmi.ac.cn

BACKGROUND: The diverse functions of ncRNAs critically depend on their

structures. Mutations in ncRNAs disrupting the structures of functional sites are

expected to be deleterious. RNA deleterious mutations have attracted wide

attentions because some of them in cells result in serious disease, and some

others in microbes influence their fitness.

RESULTS: The RDMAS web server we describe here is an online tool for evaluating

structural deleteriousness of single nucleotide mutation in RNA genes. Several

structure comparison methods have been integrated; sub-optimal structures

predicted can be optionally involved to mitigate the uncertainty of secondary

structure prediction. With a user-friendly interface, the web application is easy

to use. Intuitive illustrations are provided along with the original

computational results to facilitate quick analysis.

CONCLUSION: RDMAS can be used to explore the structure alterations which cause

mutations pathogenic, and to predict deleterious mutations which may help to

determine the functionally critical regions. RDMAS is freely accessed via

http://biosrv1.bmi.ac.cn/rdmas.

DOI: 10.1186/1471-2105-7-404

PMCID: PMC1574353

PMID: 16956394 [Indexed for MEDLINE]

2852. BMC Bioinformatics. 2006 Sep 5;7:402.

Distill: a suite of web servers for the prediction of one-, two- and

three-dimensional structural features of proteins.

Baú D(1), Martin AJ, Mooney C, Vullo A, Walsh I, Pollastri G.

Author information:

(1)School of Computer Science and Informatics, University College Dublin,

Belfield, Dublin 4, Ireland. davide.bau@ucd.ie

BACKGROUND: We describe Distill, a suite of servers for the prediction of protein

structural features: secondary structure; relative solvent accessibility; contact

density; backbone structural motifs; residue contact maps at 6, 8 and 12

Angstrom; coarse protein topology. The servers are based on large-scale ensembles

of recursive neural networks and trained on large, up-to-date, non-redundant

subsets of the Protein Data Bank. Together with structural feature predictions,

Distill includes a server for prediction of Calpha traces for short proteins (up

to 200 amino acids).

RESULTS: The servers are state-of-the-art, with secondary structure predicted

correctly for nearly 80% of residues (currently the top performance on EVA),

2-class solvent accessibility nearly 80% correct, and contact maps exceeding 50%

precision on the top non-diagonal contacts. A preliminary implementation of the

predictor of protein Calpha traces featured among the top 20 Novel Fold

predictors at the last CASP6 experiment as group Distill (ID 0348). The majority

of the servers, including the Calpha trace predictor, now take into account

homology information from the PDB, when available, resulting in greatly improved

reliability.

CONCLUSION: All predictions are freely available through a simple joint web

interface and the results are returned by email. In a single submission the user

can send protein sequences for a total of up to 32k residues to all or a

selection of the servers. Distill is accessible at the address:

http://distill.ucd.ie/distill/.

DOI: 10.1186/1471-2105-7-402

PMCID: PMC1574355

PMID: 16953874 [Indexed for MEDLINE]

2853. Biochem Biophys Res Commun. 2006 Sep 1;347(3):574-80. Epub 2006 Jun 21.

Prediction of pi-turns in proteins using PSI-BLAST profiles and secondary

structure information.

Wang Y(1), Xue ZD, Shi XH, Xu J.

Author information:

(1)Department of Control Science and Engineering, Huazhong University of Science

and Technology, Wuhan City, China. yanw@mail.hust.edu.cn

Due to the structural and functional importance of tight turns, some methods have

been proposed to predict gamma-turns, beta-turns, and alpha-turns in proteins. In

the past, studies of pi-turns were made, but not a single prediction approach has

been developed so far. It will be useful to develop a method for identifying

pi-turns in a protein sequence. In this paper, the support vector machine (SVM)

method has been introduced to predict pi-turns from the amino acid sequence. The

training and testing of this approach is performed with a newly collected data

set of 640 non-homologous protein chains containing 1931 pi-turns. Different

sequence encoding schemes have been explored in order to investigate their

effects on the prediction performance. With multiple sequence alignment and

predicted secondary structure, the final SVM model yields a Matthews correlation

coefficient (MCC) of 0.556 by a 7-fold cross-validation. A web server

implementing the prediction method is available at the following URL:

http://210.42.106.80/piturn/.

DOI: 10.1016/j.bbrc.2006.06.066

PMID: 16844090 [Indexed for MEDLINE]

2854. Bioinformatics. 2006 Sep 1;22(17):2081-6. Epub 2006 Jul 12.

An initial strategy for comparing proteins at the domain architecture level.

Lin K(1), Zhu L, Zhang DY.

Author information:

(1)MOE Key Laboratory for Biodiversity Science and Ecological Engineering and

College of Life Sciences, Beijing Normal University, Beijing 100875, China.

linkui@bnu.edu.cn

MOTIVATION: Ideally, only proteins that exhibit highly similar domain

architectures should be compared with one another as homologues or be classified

into a single family. By combining three different indices, the Jaccard index,

the Goodman-Kruskal gamma function and the domain duplicate index, into a single

similarity measure, we propose a method for comparing proteins based on their

domain architectures.

RESULTS: Evaluation of the method using the eukaryotic orthologous groups of

proteins (KOGs) database indicated that it allows the automatic and efficient

comparison of multiple-domain proteins, which are usually refractory to classic

approaches based on sequence similarity measures. As a case study, the PDZ and

LRR\_1 domains are used to demonstrate how proteins containing promiscuous domains

can be clearly compared using our method. For the convenience of users, a web

server was set up where three different query interfaces were implemented to

compare different domain architectures or proteins with domain(s), and to

identify the relationships among domain architectures within a given KOG from the

Clusters of Orthologous Groups of Proteins database.

CONCLUSION: The approach we propose is suitable for estimating the similarity of

domain architectures of proteins, especially those of multidomain proteins.

AVAILABILITY: http://cmb.bnu.edu.cn/pdart/.

DOI: 10.1093/bioinformatics/btl366

PMID: 16837531 [Indexed for MEDLINE]

2855. Bioinformatics. 2006 Sep 1;22(17):2164-5. Epub 2006 Jul 4.

MAGOS: multiple alignment and modelling server.

Garnier N(1), Friedrich A, Bolze R, Bettler E, Moulinier L, Geourjon C, Thompson

JD, Deléage G, Poch O.

Author information:

(1)Institut de Biologie et Chimie des Protéines (IBCP UMR 5086),CNRS, Univ.

Lyon1, IFR128 BioSciences Lyon-Gerland, 7, passage du Vercors, 69367 Lyon cedex

07, France.

MAGOS is a web server allowing automated protein modelling coupled to the

creation of a hierarchical and annotated multiple alignment of complete

sequences. MAGOS is designed for an interactive approach of structural

information within the framework of the evolutionary relevance of mined and

predicted sequence information.AVAILABILITY: The web server is freely available

at http://pig-pbil.ibcp.fr/magos.

DOI: 10.1093/bioinformatics/btl349

PMID: 16820425 [Indexed for MEDLINE]

2856. Curr Drug Discov Technol. 2006 Sep;3(3):167-73.

Align: a C++ class library and web server for rapid sequence alignment

prototyping.

Tosatto SC(1), Albiero A, Mantovan A, Ferrari C, Bindewald E, Toppo S.

Author information:

(1)Department of Biology, CRIBI Biotechnology Centre, University of Padova,

Padova, Italy. silvio@cribi.unipd.it

Sequence alignment remains a fundamental tool in most tasks related to the

prediction of protein sequence and structure. A C++ class library was developed

to facilitate the rapid implementation of a variety of state-of-the-art pairwise

sequence alignment techniques. These range from simple sequence to sequence to

the advanced profile to profile alignments with optional secondary structure

information. Suboptimal alignments, frequently used to estimate regions of

confidence, can also be generated. The object oriented design facilitates rapid

implementation, testing and extension of existing functionality. A simple web

interface, which can also be useful in bioinformatics education, is also

provided. Source code, online documentation and a prototypical web interface are

freely accessible to academic users from the URL:

http://protein.cribi.unipd.it/align/. A sample case study in the modelling of

human Cytochrome P450 is discussed.

PMID: 17311562 [Indexed for MEDLINE]

2857. Biochem Biophys Res Commun. 2006 Aug 18;347(1):150-7. Epub 2006 Jun 21.

Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular

localization.

Chou KC(1), Shen HB.

Author information:

(1)Gordon Life Science Institute, San Diego, CA 92130, USA. kchou@san.rr.com

Erratum in

Biochem Biophys Res Commun. 2006 Oct 6;348(4):1479.

Predicting subcellular localization of human proteins is a challenging problem,

especially when unknown query proteins do not have significant homology to

proteins of known subcellular locations and when more locations need to be

covered. To tackle the challenge, protein samples are expressed by hybridizing

the gene ontology (GO) database and amphiphilic pseudo amino acid composition

(PseAA). Based on such a representation frame, a novel ensemble classifier,

called "Hum-PLoc", was developed by fusing many basic individual classifiers

through a voting system. The "engine" of these basic classifiers was operated by

the KNN (K-nearest neighbor) rule. As a demonstration, tests were performed with

the ensemble classifier for human proteins among the following 12 locations: (1)

centriole; (2) cytoplasm; (3) cytoskeleton; (4) endoplasmic reticulum; (5)

extracell; (6) Golgi apparatus; (7) lysosome; (8) microsome; (9) mitochondrion;

(10) nucleus; (11) peroxisome; (12) plasma membrane. To get rid of redundancy and

homology bias, none of the proteins investigated here had > or = 25% sequence

identity to any other in a same subcellular location. The overall success rates

thus obtained via the jackknife cross-validation test and independent dataset

test were 81.1% and 85.0%, respectively, which are more than 50% higher than

those obtained by the other existing methods on the same stringent datasets.

Furthermore, an incisive and compelling analysis was given to elucidate that the

overwhelmingly high success rate obtained by the new predictor is by no means due

to a trivial utilization of the GO annotations. This is because, for those

proteins with "subcellular location unknown" annotation in Swiss-Prot database,

most (more than 99%) of their corresponding GO numbers in GO database are also

annotated with "cellular component unknown". The information and clues for

predicting subcellular locations of proteins are actually buried into a series of

tedious GO numbers, just like they are buried into a pile of complicated amino

acid sequences although with a different manner and "depth". To dig out the

knowledge about their locations, a sophisticated operation engine is needed. And

the current predictor is one of these kinds, and has proved to be a very powerful

one. The Hum-PLoc classifier is available as a web-server at

http://202.120.37.186/bioinf/hum.

DOI: 10.1016/j.bbrc.2006.06.059

PMID: 16808903 [Indexed for MEDLINE]

2858. BMC Bioinformatics. 2006 Aug 7;7:371.

UniFrac--an online tool for comparing microbial community diversity in a

phylogenetic context.

Lozupone C(1), Hamady M, Knight R.

Author information:

(1)Department of Molecular, Cellular, and Developmental Biology, University of

Colorado, Boulder, CO 80309, USA. lozupone@colorado.edu

BACKGROUND: Moving beyond pairwise significance tests to compare many microbial

communities simultaneously is critical for understanding large-scale trends in

microbial ecology and community assembly. Techniques that allow microbial

communities to be compared in a phylogenetic context are rapidly gaining

acceptance, but the widespread application of these techniques has been hindered

by the difficulty of performing the analyses.

RESULTS: We introduce UniFrac, a web application available at

http://bmf.colorado.edu/unifrac, that allows several phylogenetic tests for

differences among communities to be easily applied and interpreted. We

demonstrate the use of UniFrac to cluster multiple environments, and to test

which environments are significantly different. We show that analysis of

previously published sequences from the Columbia river, its estuary, and the

adjacent coastal ocean using the UniFrac interface provided insights that were

not apparent from the initial data analysis, which used other commonly employed

techniques to compare the communities.

CONCLUSION: UniFrac provides easy access to powerful multivariate techniques for

comparing microbial communities in a phylogenetic context. We thus expect that it

will provide a completely new picture of many microbial interactions and

processes in both environmental and medical contexts.

DOI: 10.1186/1471-2105-7-371

PMCID: PMC1564154

PMID: 16893466 [Indexed for MEDLINE]

2859. Immunogenetics. 2006 Aug;58(8):607-13. Epub 2006 Jul 11.

MHC-BPS: MHC-binder prediction server for identifying peptides of flexible

lengths from sequence-derived physicochemical properties.

Cui J(1), Han LY, Lin HH, Tang ZQ, Jiang L, Cao ZW, Chen YZ.

Author information:

(1)Bioinformatics and Drug Design Group, Department of Pharmacy, National

University of Singapore, Singapore, 117543, Singapore.

Major histocompatibility complex (MHC)-binding peptides are essential for antigen

recognition by T-cell receptors and are being explored for vaccine design.

Computational methods have been developed for predicting MHC-binding peptides of

fixed lengths, based on the training of relatively few non-binders. It is

desirable to introduce methods applicable for peptides of flexible lengths and

trained by using more diverse sets of non-binders. MHC-BPS is a web-based

MHC-binder prediction server that uses support vector machines for predicting

peptide binders of flexible lengths for 18 MHC class I and 12 class II alleles

from sequence-derived physicochemical properties, which were trained by using

4,208 approximately 3,252 binders and 234,333 approximately 168,793 non-binders,

and evaluated by an independent set of 545 approximately 476 binders and 110,564

approximately 84,430 non-binders. The binder prediction accuracies are 86

approximately 99% for 25 and 70 approximately 80% for five alleles, and the

non-binder accuracies are 96 approximately 99% for 30 alleles. A screening of

HIV-1 genome identifies 0.01 approximately 5% and 5 approximately 8% of the

constituent peptides as binders for 24 and 6 alleles, respectively, including 75

approximately 100% of the known epitopes. This method correctly predicts 73.3% of

the 15 newly published epitopes in the last 4 months of 2005. MHC-BPS is

available at http://bidd.cz3.nus.edu.sg/mhc/ .

DOI: 10.1007/s00251-006-0117-2

PMID: 16832638 [Indexed for MEDLINE]

2860. J Proteome Res. 2006 Aug;5(8):1888-97.

Predicting eukaryotic protein subcellular location by fusing optimized

evidence-theoretic K-Nearest Neighbor classifiers.

Chou KC(1), Shen HB.

Author information:

(1)Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego,

California 92130, USA. kchou@san.rr.com

Facing the explosion of newly generated protein sequences in the post genomic

era, we are challenged to develop an automated method for fast and reliably

annotating their subcellular locations. Knowledge of subcellular locations of

proteins can provide useful hints for revealing their functions and understanding

how they interact with each other in cellular networking. Unfortunately, it is

both expensive and time-consuming to determine the localization of an

uncharacterized protein in a living cell purely based on experiments. To tackle

the challenge, a novel hybridization classifier was developed by fusing many

basic individual classifiers through a voting system. The "engine" of these basic

classifiers was operated by the OET-KNN (Optimized Evidence-Theoretic K-Nearest

Neighbor) rule. As a demonstration, predictions were performed with the fusion

classifier for proteins among the following 16 localizations: (1) cell wall, (2)

centriole, (3) chloroplast, (4) cyanelle, (5) cytoplasm, (6) cytoskeleton, (7)

endoplasmic reticulum, (8) extracell, (9) Golgi apparatus, (10) lysosome, (11)

mitochondria, (12) nucleus, (13) peroxisome, (14) plasma membrane, (15) plastid,

and (16) vacuole. To get rid of redundancy and homology bias, none of the

proteins investigated here had >/=25% sequence identity to any other in a same

subcellular location. The overall success rates thus obtained via the jack-knife

cross-validation test and independent dataset test were 81.6% and 83.7%,

respectively, which were 46 approximately 63% higher than those performed by the

other existing methods on the same benchmark datasets. Also, it is clearly

elucidated that the overwhelmingly high success rates obtained by the fusion

classifier is by no means a trivial utilization of the GO annotations as prone to

be misinterpreted because there is a huge number of proteins with given accession

numbers and the corresponding GO numbers, but their subcellular locations are

still unknown, and that the percentage of proteins with GO annotations indicating

their subcellular components is even less than the percentage of proteins with

known subcellular location annotation in the Swiss-Prot database. It is

anticipated that the powerful fusion classifier may also become a very useful

high throughput tool in characterizing other attributes of proteins according to

their sequences, such as enzyme class, membrane protein type, and nuclear

receptor subfamily, among many others. A web server, called "Euk-OET-PLoc", has

been designed at http://202.120.37.186/bioinf/euk-oet for public to predict

subcellular locations of eukaryotic proteins by the fusion OET-KNN classifier.

DOI: 10.1021/pr060167c

PMID: 16889410 [Indexed for MEDLINE]

2861. Bioinformation. 2006 Jul 19;1(5):176-9.

LIPPRED: A web server for accurate prediction of lipoprotein signal sequences and

cleavage sites.

Taylor PD(1), Toseland CP, Attwood TK, Flower DR.

Author information:

(1)The Jenner Institute, University of Oxford, Compton, Newbury, Berkshire, RG20

7NN, UK.

Bacterial lipoproteins have many important functions and represent a class of

possible vaccine candidates. The prediction of lipoproteins from sequence is thus

an important task for computational vaccinology. Naïve-Bayesian networks were

trained to identify SpaseII cleavage sites and their preceding signal sequences

using a set of 199 distinct lipoprotein sequences. A comprehensive range of

sequence models was used to identify the best model for lipoprotein signal

sequences. The best performing sequence model was found to be 10-residues in

length, including the conserved cysteine lipid attachment site and the nine

residues prior to it. The sensitivity of prediction for LipPred was 0.979, while

the specificity was 0.742. Here, we describe LipPred, a web server for

lipoprotein prediction; available at the URL: http://www.jenner.ac.uk/LipPred/.

LipPred is the most accurate method available for the detection of SpaseIIcleaved

lipoprotein signal sequences and the prediction of their cleavage sites.

PMCID: PMC1891677

PMID: 17597883

2862. Bioinformatics. 2006 Jul 15;22(14):1784-5. Epub 2006 Jun 20.

TSSub: eukaryotic protein subcellular localization by extracting features from

profiles.

Guo J(1), Lin Y.

Author information:

(1)Laboratory of Statistical Computation & Bioinformatics, Department of

Mathematical Sciences, Tsinghua University Beijing 100084, China.

guojian99@tsinghua.org.cn

This paper introduces a new subcellular localization system (TSSub) for

eukaryotic proteins. This system extracts features from both profiles and amino

acid sequences. Four different features are extracted from profiles by four

probabilistic neural network (PNN) classifiers, respectively (the amino acid

composition from whole profiles; the amino acid composition from the N-terminus

of profiles; the dipeptide composition from whole profiles and the amino acid

composition from fragments of profiles). In addition, a support vector machine

(SVM) classifier is added to implement the residue-couple feature extracted from

amino acid sequences. The results from the five classifiers are fused by an

additional SVM classifier. The overall accuracies of this TSSub reach 93.0 and

77.4% on Reinhardt and Hubbard's eukaryotic protein dataset and Huang and Li's

eukaryotic protein dataset, respectively. The comparison with existing methods

results shows TSSub provides better prediction performance than existing

methods.AVAILABILITY: The web server is available from

http://166.111.24.5/webtools/TSSub/index.html.

DOI: 10.1093/bioinformatics/btl180

PMID: 16787975 [Indexed for MEDLINE]

2863. Bioinformatics. 2006 Jul 15;22(14):1788-9. Epub 2006 May 18.

TRFMA: a web-based tool for terminal restriction fragment length polymorphism

analysis based on molecular weight.

Nakano Y(1), Takeshita T, Yamashita Y.

Author information:

(1)Department of Preventive Dentistry, Faculty of Dental Science, Kyushu

University, Fukuoka-shi, Japan. yosh@dent.kyushu-u.ac.jp

TRFMA provides a Web environment for analyzing T-RFLP results based on molecular

weights of the fragments, rather than the numbers of nucleotides, to increase

accuracy. The 16S rRNA data are saved as an XML file containing around 650

sequences (light version) and a MySQL database containing around 50 000 sequences

(full version), which are connected to Web server via PHP5 and manipulated on an

Internet browser.AVAILABILITY: TRFMA is freely available at

http://myamagu.dent.kyushu-u.ac.jp/bioinformatics/trfma/index.html and can be

downloaded from the same site.

DOI: 10.1093/bioinformatics/btl186

PMID: 16709590 [Indexed for MEDLINE]

2864. Bioinformatics. 2006 Jul 15;22(14):1723-9. Epub 2006 May 11.

SCARNA: fast and accurate structural alignment of RNA sequences by matching

fixed-length stem fragments.

Tabei Y(1), Tsuda K, Kin T, Asai K.

Author information:

(1)Department of Computational Biology, Graduate School of Frontier Science,

University of Tokyo, CB04 Kiban-tou 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8561,

Japan. scarna@m.aist.go.jp

MOTIVATION: The functions of non-coding RNAs are strongly related to their

secondary structures, but it is known that a secondary structure prediction of a

single sequence is not reliable. Therefore, we have to collect similar RNA

sequences with a common secondary structure for the analyses of a new non-coding

RNA without knowing the exact secondary structure itself. Therefore, the sequence

comparison in searching similar RNAs should consider not only their sequence

similarities but also their potential secondary structures. Sankoff's algorithm

predicts the common secondary structures of the sequences, but it is

computationally too expensive to apply to large-scale analyses. Because we often

want to compare a large number of cDNA sequences or to search similar RNAs in the

whole genome sequences, much faster algorithms are required.

RESULTS: We propose a new method of comparing RNA sequences based on the

structural alignments of the fixed-length fragments of the stem candidates. The

implemented software, SCARNA (Stem Candidate Aligner for RNAs), is fast enough to

apply to the long sequences in the large-scale analyses. The accuracy of the

alignments is better or comparable with the much slower existing algorithms.

AVAILABILITY: The web server of SCARNA with graphical structural alignment viewer

is available at http://www.scarna.org/.

DOI: 10.1093/bioinformatics/btl177

PMID: 16690634 [Indexed for MEDLINE]

2865. Bioinformatics. 2006 Jul 15;22(14):1717-22. Epub 2006 May 3.

Ensemble classifier for protein fold pattern recognition.

Shen HB(1), Chou KC.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, Shanghai 200030, China. lifesci-sjtu@san.rr.com

MOTIVATION: Prediction of protein folding patterns is one level deeper than that

of protein structural classes, and hence is much more complicated and difficult.

To deal with such a challenging problem, the ensemble classifier was introduced.

It was formed by a set of basic classifiers, with each trained in different

parameter systems, such as predicted secondary structure, hydrophobicity, van der

Waals volume, polarity, polarizability, as well as different dimensions of

pseudo-amino acid composition, which were extracted from a training dataset. The

operation engine for the constituent individual classifiers was OET-KNN

(optimized evidence-theoretic k-nearest neighbors) rule. Their outcomes were

combined through a weighted voting to give a final determination for classifying

a query protein. The recognition was to find the true fold among the 27 possible

patterns.

RESULTS: The overall success rate thus obtained was 62% for a testing dataset

where most of the proteins have <25% sequence identity with the proteins used in

training the classifier. Such a rate is 6-21% higher than the corresponding rates

obtained by various existing NN (neural networks) and SVM (support vector

machines) approaches, implying that the ensemble classifier is very promising and

might become a useful vehicle in protein science, as well as proteomics and

bioinformatics.

AVAILABILITY: The ensemble classifier, called PFP-Pred, is available as a

web-server at http://202.120.37.186/bioinf/fold/PFP-Pred.htm for public usage.

DOI: 10.1093/bioinformatics/btl170

PMID: 16672258 [Indexed for MEDLINE]

2866. Bioinformatics. 2006 Jul 15;22(14):1794-5. Epub 2006 May 3.

SHARP2: protein-protein interaction predictions using patch analysis.

Murakami Y(1), Jones S.

Author information:

(1)Department of Biochemistry, School of Life Sciences, John Maynard Smith

Building, University of Sussex, Falmer, Brighton, BN1 9QG, UK.

SHARP2 is a flexible web-based bioinformatics tool for predicting potential

protein-protein interaction sites on protein structures. It implements a

predictive algorithm that calculates multiple parameters for overlapping patches

of residues on the surface of a protein. Six parameters are calculated: solvation

potential, hydrophobicity, accessible surface area, residue interface propensity,

planarity and protrusion (SHARP2). Parameter scores for each patch are combined,

and the patch with the highest combined score is predicted as a potential

interaction site. SHARP2 enables users to upload 3D protein structure files in

PDB format, to obtain information on potential interaction sites as downloadable

HTML tables and to view the location of the sites on the 3D structure using Jmol.

The server allows for the input of multiple structures and multiple combinations

of parameters. Therefore predictions can be made for complete datasets, as well

as individual structures.AVAILABILITY:

http://www.bioinformatics.sussex.ac.uk/SHARP2.

DOI: 10.1093/bioinformatics/btl171

PMID: 16672257 [Indexed for MEDLINE]

2867. Bioinformatics. 2006 Jul 15;22(14):1690-701. Epub 2006 Apr 13.

Ribosomal RNA as molecular barcodes: a simple correlation analysis without

sequence alignment.

Chu KH(1), Li CP, Qi J.

Author information:

(1)Department of Biology, The Chinese University of Hong Kong, Hong Kong, China.

kahouchu@cuhk.edu.hk

MOTIVATION: We explored the feasibility of using unaligned rRNA gene sequences as

DNA barcodes, based on correlation analysis of composition vectors (CVs) derived

from nucleotide strings. We tested this method with seven rRNA (including 12, 16,

18, 26 and 28S) datasets from a wide variety of organisms (from archaea to

tetrapods) at taxonomic levels ranging from class to species.

RESULT: Our results indicate that grouping of taxa based on CV analysis is always

in good agreement with the phylogenetic trees generated by traditional

approaches, although in some cases the relationships among the higher systemic

groups may differ. The effectiveness of our analysis might be related to the

length and divergence among sequences in a dataset. Nevertheless, the correct

grouping of sequences and accurate assignment of unknown taxa make our analysis a

reliable and convenient approach in analyzing unaligned sequence datasets of

various rRNAs for barcoding purposes.

AVAILABILITY: The newly designed software (CVTree 1.0) is publicly available at

the Composition Vector Tree (CVTree) web server http://cvtree.cbi.pku.edu.cn.

DOI: 10.1093/bioinformatics/btl146

PMID: 16613905 [Indexed for MEDLINE]

2868. J Comput Aided Mol Des. 2006 Jul-Aug;20(7-8):519-27. Epub 2006 Nov 11.

Milestones in electron crystallography.

Renault L(1), Chou HT, Chiu PL, Hill RM, Zeng X, Gipson B, Zhang ZY, Cheng A,

Unger V, Stahlberg H.

Author information:

(1)Molecular and Cellular Biology, College of Biological Sciences, University of

California at Davis, Briggs Hall, 1 Shields Ave., Davis, CA 95616, USA.

Electron crystallography determines the structure of membrane embedded proteins

in the two-dimensionally crystallized state by cryo-transmission electron

microscopy imaging and computer structure reconstruction. Milestones on the path

to the structure are high-level expression, purification of functional protein,

reconstitution into two-dimensional lipid membrane crystals, high-resolution

imaging, and structure determination by computer image processing. Here we review

the current state of these methods. We also created an Internet information

exchange platform for electron crystallography, where guidelines for imaging and

data processing method are maintained. The server (http://2dx.org) provides the

electron crystallography community with a central information exchange platform,

which is structured in blog and Wiki form, allowing visitors to add comments or

discussions. It currently offers a detailed step-by-step introduction to image

processing with the MRC software program. The server is also a repository for the

2dx software package, a user-friendly image processing system for 2D membrane

protein crystals.

DOI: 10.1007/s10822-006-9075-x

PMCID: PMC2194810

PMID: 17103018 [Indexed for MEDLINE]

2869. Magn Reson Chem. 2006 Jul;44 Spec No:S158-67.

Accurate prediction of protein torsion angles using chemical shifts and sequence

homology.

Neal S(1), Berjanskii M, Zhang H, Wishart DS.

Author information:

(1)Department of Biological Sciences, University of Alberta, Edmonton, AB,

Canada.

Torsion angle restraints are frequently used in the determination and refinement

of protein structures by NMR. These restraints may be obtained by J coupling,

cross-correlation measurements, nuclear Overhauser effects (NOEs) or secondary

chemical shifts. Currently most backbone (phi/psi) torsion angles are determined

using a combination of J(HNHalpha) couplings and chemical shift measurements

while most side-chain (chi1) angles and cis/trans peptide bond angles (omega) are

determined via NOEs. The dependency on multiple experimental (and computational)

methods to obtain different torsion angle restraints is both time-consuming and

error prone. The situation could be greatly improved if the determination of all

torsion angles (phi, psi, chi and omega) could be made via a single type of

measurement (i.e. chemical shifts). Here we describe a program, called SHIFTOR,

that is able to accurately predict a large number of protein torsion angles (phi,

psi, omega, chi1) using only 1H, 13C and 15N chemical shift assignments as input.

Overall, the program is 100x faster and its predictions are approximately 20%

better than existing methods. The program is also capable of predicting chi1

angles with 81% accuracy and omega angles with 100% accuracy. SHIFTOR exploits

many of the recent developments and observations regarding chemical shift

dependencies as well as using information in the Protein Databank to improve the

quality of its shift-derived torsion angle predictions. SHIFTOR is available as a

freely accessible web server at http://wishart.biology.ualberta.ca/shiftor.

Copyright 2006 John Wiley & Sons, Ltd.

DOI: 10.1002/mrc.1832

PMID: 16823900 [Indexed for MEDLINE]

2870. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W99-103.

Localizome: a server for identifying transmembrane topologies and TM helices of

eukaryotic proteins utilizing domain information.

Lee S(1), Lee B, Jang I, Kim S, Bhak J.

Author information:

(1)National Genome Information Center, Korea Research Institute of Bioscience and

Biotechnology, 52 Ueun-dong, Yuseong-gu, Daejeon 305-333, Republic of Korea.

The Localizome server predicts the transmembrane (TM) helix number and TM

topology of a user-supplied eukaryotic protein and presents the result as an

intuitive graphic representation. It utilizes hmmpfam to detect the presence of

Pfam domains and a prediction algorithm, Phobius, to predict the TM helices. The

results are combined and checked against the TM topology rules stored in a

protein domain database called LocaloDom. LocaloDom is a curated database that

contains TM topologies and TM helix numbers of known protein domains. It was

constructed from Pfam domains combined with Swiss-Prot annotations and Phobius

predictions. The Localizome server corrects the combined results of the user

sequence to conform to the rules stored in LocaloDom. Compared with other

programs, this server showed the highest accuracy for TM topology prediction: for

soluble proteins, the accuracy and coverage were 99 and 75%, respectively, while

for TM protein domain regions, they were 96 and 68%, respectively. With a

graphical representation of TM topology and TM helix positions with the domain

units, the Localizome server is a highly accurate and comprehensive information

source for subcellular localization for soluble proteins as well as membrane

proteins. The Localizome server can be found at http://localizome.org/.

DOI: 10.1093/nar/gkl351

PMCID: PMC1538878

PMID: 16845118 [Indexed for MEDLINE]

2871. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W95-8.

OPAAS: a web server for optimal, permuted, and other alternative alignments of

protein structures.

Shih ES(1), Gan RC, Hwang MJ.

Author information:

(1)Institute of Biomedical Sciences, Academia Sinica, Taipei 11529, Taiwan.

The large number of experimentally determined protein 3D structures is a rich

resource for studying protein function and evolution, and protein structure

comparison (PSC) is a key method for such studies. When comparing two protein

structures, almost all currently available PSC servers report a single and

sequential (i.e. topological) alignment, whereas the existence of good

alternative alignments, including those involving permutations (i.e.

non-sequential or non-topological alignments), is well known. We have recently

developed a novel PSC method that can detect alternative alignments of

statistical significance (alignment similarity P-value <10(-5)), including

structural permutations at all levels of complexity. OPAAS, the server of this

PSC method freely accessible at our website (http://opaas.ibms.sinica.edu.tw),

provides an easy-to-read hierarchical layout of output to display detailed

information on all of the significant alternative alignments detected. Because

these alternative alignments can offer a more complete picture on the structural,

evolutionary and functional relationship between two proteins, OPAAS can be used

in structural bioinformatics research to gain additional insight that is not

readily provided by existing PSC servers.

DOI: 10.1093/nar/gkl264

PMCID: PMC1538888

PMID: 16845117 [Indexed for MEDLINE]

2872. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W79-83.

ProSAT2--Protein Structure Annotation Server.

Gabdoulline RR(1), Ulbrich S, Richter S, Wade RC.

Author information:

(1)Molecular and Cellular Modeling Group, EML Research, Schloss-Wolfsbrunnenweg

33, 69118 Heidelberg, Germany. razif.gabdoulline@eml-r.villa-bosch.de

ProSAT2 is a server to facilitate interactive visualization of sequence-based,

residue-specific annotations mapped onto 3D protein structures. As the successor

of ProSAT (Protein Structure Annotation Tool), it includes its features for

visualizing SwissProt and PROSITE functional annotations. Currently, the ProSAT2

server can perform automated mapping of information on variants and mutations

from the UniProt KnowledgeBase and the BRENDA enzyme information system onto

protein structures. It also accepts and maps user-prepared annotations. By means

of an annotation selector, the user can interactively select and group

residue-based information according to criteria such as whether a mutation

affects enzyme activity. The visualization of the protein structures is based on

the WebMol Java molecular viewer and permits simultaneous highlighting of

annotated residues and viewing of the corresponding descriptive texts. ProSAT2 is

available at http://projects.villa-bosch.de/mcm/database/prosat2/.

DOI: 10.1093/nar/gkl216

PMCID: PMC1538895

PMID: 16845114 [Indexed for MEDLINE]

2873. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W75-8.

Protein Peeling 2: a web server to convert protein structures into series of

protein units.

Gelly JC(1), Etchebest C, Hazout S, de Brevern AG.

Author information:

(1)CNRS, UPR 9080, Laboratoire de Biochimie Théorique, Institut de Biologie,

Physico-Chimique (FRC 550) 11, rue Pierre et Marie Curie, 75005 Paris, France.

Protein Peeling 2 (PP2) is a web server for the automatic identification of

protein units (PUs) given the 3D coordinates of a protein. PUs are an

intermediate level of protein structure description between protein domains and

secondary structures. It is a new tool to better understand and analyze the

organization of protein structures. PP2 uses only the matrices of protein contact

probabilities and cuts the protein structures optimally using Matthews'

coefficient correlation. An index assesses the compactness quality of each PU.

Results are given both textually and graphically using JMol and PyMol softwares.

The server can be accessed from http://www.ebgm.jussieu.fr/~gelly/index.html.

DOI: 10.1093/nar/gkl292

PMCID: PMC1538916

PMID: 16845113 [Indexed for MEDLINE]

2874. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W725-8.

metaSHARK: a WWW platform for interactive exploration of metabolic networks.

Hyland C(1), Pinney JW, McConkey GA, Westhead DR.

Author information:

(1)Faculty of Biological Sciences, University of Leeds, Clarendon Way, Leeds LS2

9JT, UK.

The metaSHARK (metabolic search and reconstruction kit) web server offers users

an intuitive, fully interactive way to explore the KEGG metabolic network via a

WWW browser. Metabolic reconstruction information for specific organisms,

produced by our automated SHARKhunt tool or from other programs or genome

annotations, may be uploaded to the website and overlaid on the generic network.

Additional data from gene expression experiments can also be incorporated,

allowing the visualization of differential gene expression in the context of the

predicted metabolic network. metaSHARK is available at

http://bioinformatics.leeds.ac.uk/shark/.

DOI: 10.1093/nar/gkl196

PMCID: PMC1538829

PMID: 16845107 [Indexed for MEDLINE]

2875. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W720-4.

KOBAS server: a web-based platform for automated annotation and pathway

identification.

Wu J(1), Mao X, Cai T, Luo J, Wei L.

Author information:

(1)Center for Bioinformatics, National Laboratory of Protein Engineering and

Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing

100871, PR China.

There is an increasing need to automatically annotate a set of genes or proteins

(from genome sequencing, DNA microarray analysis or protein 2D gel experiments)

using controlled vocabularies and identify the pathways involved, especially the

statistically enriched pathways. We have previously demonstrated the KEGG

Orthology (KO) as an effective alternative controlled vocabulary and developed a

standalone KO-Based Annotation System (KOBAS). Here we report a KOBAS server with

a friendly web-based user interface and enhanced functionalities. The server can

support input by nucleotide or amino acid sequences or by sequence identifiers in

popular databases and can annotate the input with KO terms and KEGG pathways by

BLAST sequence similarity or directly ID mapping to genes with known annotations.

The server can then identify both frequent and statistically enriched pathways,

offering the choices of four statistical tests and the option of multiple testing

correction. The server also has a 'User Space' in which frequent users may store

and manage their data and results online. We demonstrate the usability of the

server by finding statistically enriched pathways in a set of upregulated genes

in Alzheimer's Disease (AD) hippocampal cornu ammonis 1 (CA1). KOBAS server can

be accessed at http://kobas.cbi.pku.edu.cn.

DOI: 10.1093/nar/gkl167

PMCID: PMC1538915

PMID: 16845106 [Indexed for MEDLINE]

2876. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W714-9.

The Path-A metabolic pathway prediction web server.

Pireddu L(1), Szafron D, Lu P, Greiner R.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB, Canada

T6G 2E8.

Pathway Analyst (Path-A) is a publicly available web server

(http://path-a.cs.ualberta.ca) that predicts metabolic pathways. It takes a FASTA

format file containing a set of query protein sequences from a single organism (a

partial or complete proteome) and identifies those sequences that are likely to

participate in any of its supported metabolic pathways (currently 10). Path-A

uses a number of machine-learning and sequence analysis techniques (e.g. SVM,

BLAST and HMM) to predict pathways. Each machine-learned classifier exploits

similarity between sequences in the pathways of its model organisms and sequences

in the query set. It predicts the pathways that are present in the query organism

and annotates each predicted reaction and catalyst, using the appropriate

sequences from the query set. Path-A also provides a browsable and searchable

database of the pathways for the model organisms that are used to make its

predictions. Path-A's predictor sets (using different classifier technologies)

have been evaluated using standard cross-validation techniques on a dataset of 10

metabolic pathways across 13 model organisms--a total of 125 organism-specific

pathways. The most accurate classifier technology obtained a mean precision of

78.3% and a mean recall of 92.6% in predicting all catalyst proteins, of all

reactions, in all pathways present in the dataset. Although Path-A currently only

supports metabolic pathways, the underlying prediction techniques are general

enough for other types of pathways. Consequently, it is our intent to extend

Path-A to predict other types of pathways, including signalling pathways.

DOI: 10.1093/nar/gkl228

PMCID: PMC1538809

PMID: 16845105 [Indexed for MEDLINE]

2877. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W704-7.

The internal transcribed spacer 2 database--a web server for (not only) low level

phylogenetic analyses.

Schultz J(1), Müller T, Achtziger M, Seibel PN, Dandekar T, Wolf M.

Author information:

(1)Department of Bioinformatics, Biocenter, University of Würzburg Am Hubland,

97074 Würzburg, Germany. Joerg.Schultz@biozentrum.uni-wuerzburg.de

The internal transcribed spacer 2 (ITS2) is a phylogenetic marker which has been

of broad use in generic and infrageneric level classifications, as its sequence

evolves comparably fast. Only recently, it became clear, that the ITS2 might be

useful even for higher level systematic analyses. As the secondary structure is

highly conserved within all eukaryotes it serves as a valuable template for the

construction of highly reliable sequence-structure alignments, which build a

fundament for subsequent analyses. Thus, any phylogenetic study using ITS2 has to

consider both sequence and structure. We have integrated a homology based RNA

structure prediction algorithm into a web server, which allows the detection and

secondary structure prediction for ITS2 in any given sequence. Furthermore, the

resource contains more than 25,000 pre-calculated secondary structures for the

currently known ITS2 sequences. These can be taxonomically searched and browsed.

Thus, our resource could become a starting point for ITS2-based phylogenetic

analyses and is therefore complementary to databases of other phylogenetic

markers, which focus on higher level analyses. The current version of the ITS2

database can be accessed via http://its2.bioapps.biozentrum.uni-wuerzburg.de.

DOI: 10.1093/nar/gkl129

PMCID: PMC1538906

PMID: 16845103 [Indexed for MEDLINE]

2878. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W700-3.

ModelTest Server: a web-based tool for the statistical selection of models of

nucleotide substitution online.

Posada D(1).

Author information:

(1)Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, 36310

Vigo, Spain. dposada@uvigo.es

ModelTest server is a web-based application for the selection of models of

nucleotide substitution using the program ModelTest. The server takes as input a

text file with likelihood scores for the set of candidate models. Models can be

selected with hierarchical likelihood ratio tests, or with the Akaike or Bayesian

information criteria. The output includes several statistics for the assessment

of model selection uncertainty, for model averaging or to estimate the relative

importance of model parameters. The server can be accessed at

http://darwin.uvigo.es/software/modeltest\_server.html.

DOI: 10.1093/nar/gkl042

PMCID: PMC1538795

PMID: 16845102 [Indexed for MEDLINE]

2879. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W70-4.

FOLD-RATE: prediction of protein folding rates from amino acid sequence.

Gromiha MM(1), Thangakani AM, Selvaraj S.

Author information:

(1)Computational Biology Research Center, National Institute of Advanced

Industrial Science and Technology, AIST Tokyo Waterfront Bio-IT Research

Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan. michael-gromiha@aist.go.jp

We have developed a web server, FOLD-RATE, for predicting the folding rates of

proteins from their amino acid sequences. The relationship between amino acid

properties and protein folding rates has been systematically analyzed and a

statistical method based on linear regression technique has been proposed for

predicting the folding rate of proteins. We found that the classification of

proteins into different structural classes shows an excellent correlation between

amino acid properties and folding rates of two and three-state proteins.

Consequently, different regression equations have been developed for proteins

belonging to all-alpha, all-beta and mixed class. We observed an excellent

agreement between predicted and experimentally observed folding rates of

proteins; the correlation coefficients are, 0.99, 0.97 and 0.90, respectively,

for all-alpha, all-beta and mixed class proteins. The prediction server is freely

available at http://psfs.cbrc.jp/fold-rate/.

DOI: 10.1093/nar/gkl043

PMCID: PMC1538837

PMID: 16845101 [Indexed for MEDLINE]

2880. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W692-5.

CorGen--measuring and generating long-range correlations for DNA sequence

analysis.

Messer PW(1), Arndt PF.

Author information:

(1)Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin,

Germany. philipp.messer@molgen.mpg.de

CorGen is a web server that measures long-range correlations in the base

composition of DNA and generates random sequences with the same correlation

parameters. Long-range correlations are characterized by a power-law decay of the

auto correlation function of the GC-content. The widespread presence of such

correlations in eukaryotic genomes calls for their incorporation into accurate

null models of eukaryotic DNA in computational biology. For example, the score

statistics of sequence alignment and the performance of motif finding algorithms

are significantly affected by the presence of genomic long-range correlations. We

use an expansion-randomization dynamics to efficiently generate the correlated

random sequences. The server is available at http://corgen.molgen.mpg.de.

DOI: 10.1093/nar/gkl234

PMCID: PMC1538783

PMID: 16845099 [Indexed for MEDLINE]

2881. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W683-5.

Quadfinder: server for identification and analysis of quadruplex-forming motifs

in nucleotide sequences.

Scaria V(1), Hariharan M, Arora A, Maiti S.

Author information:

(1)GN Ramachandran Knowledge Center for Genome Informatics, Institute of Genomics

and Integrative Biology CSIR, Mall Road, Delhi 110 007, India.

G-quadruplex secondary structures, which play a structural role in repetitive DNA

such as telomeres, may also play a functional role at other genomic locations as

targetable regulatory elements which control gene expression. The recent interest

in application of quadruplexes in biological systems prompted us to develop a

tool for the identification and analysis of quadruplex-forming nucleotide

sequences especially in the RNA. Here we present Quadfinder, an online server for

prediction and bioinformatics of uni-molecular quadruplex-forming nucleotide

sequences. The server is designed to be user-friendly and needs minimal

intervention by the user, while providing flexibility of defining the variants of

the motif. The server is freely available at URL

http://miracle.igib.res.in/quadfinder/.

DOI: 10.1093/nar/gkl299

PMCID: PMC1538896

PMID: 16845097 [Indexed for MEDLINE]

2882. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W676-82.

QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide

sequences.

Kikin O(1), D'Antonio L, Bagga PS.

Author information:

(1)Bioinformatics, School of Theoretical and Applied Science, Ramapo College of

New Jersey, Mahwah, NJ 07430, USA.

The quadruplex structures formed by guanine-rich nucleic acid sequences have

received significant attention recently because of growing evidence for their

role in important biological processes and as therapeutic targets. G-quadruplex

DNA has been suggested to regulate DNA replication and may control cellular

proliferation. Sequences capable of forming G-quadruplexes in the RNA have been

shown to play significant roles in regulation of polyadenylation and splicing

events in mammalian transcripts. Whether quadruplex structure directly plays a

role in regulating RNA processing requires investigation. Computational

approaches to study G-quadruplexes allow detailed analysis of mammalian genomes.

There are no known easily accessible user-friendly tools that can compute

G-quadruplexes in the nucleotide sequences. We have developed a web-based server,

QGRS Mapper, that predicts quadruplex forming G-rich sequences (QGRS) in

nucleotide sequences. It is a user-friendly application that provides many

options for defining and studying G-quadruplexes. It performs analysis of the

user provided genomic sequences, e.g. promoter and telomeric regions, as well as

RNA sequences. It is also useful for predicting G-quadruplex structures in

oligonucleotides. The program provides options to search and retrieve desired

gene/nucleotide sequence entries from NCBI databases for mapping G-quadruplexes

in the context of RNA processing sites. This feature is very useful for

investigating the functional relevance of G-quadruplex structure, in particular

its role in regulating the gene expression by alternative processing. In addition

to providing data on composition and locations of QGRS relative to the processing

sites in the pre-mRNA sequence, QGRS Mapper features interactive graphic

representation of the data. The user can also use the graphics module to

visualize QGRS distribution patterns among all the alternative RNA products of a

gene simultaneously on a single screen. QGRS Mapper can be accessed at

http://bioinformatics.ramapo.edu/QGRS/.

DOI: 10.1093/nar/gkl253

PMCID: PMC1538864

PMID: 16845096 [Indexed for MEDLINE]

2883. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W665-9.

PrimerStation: a highly specific multiplex genomic PCR primer design server for

the human genome.

Yamada T(1), Soma H, Morishita S.

Author information:

(1)Department of Computational Biology, Graduate School of Frontier Sciences,

University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan.

yamada@cb.k.u-tokyo.ac.jp

PrimerStation (http://ps.cb.k.u-tokyo.ac.jp) is a web service that calculates

primer sets guaranteeing high specificity against the entire human genome. To

achieve high accuracy, we used the hybridization ratio of primers in liquid

solution. Calculating the status of sequence hybridization in terms of the

stringent hybridization ratio is computationally costly, and no web service

checks the entire human genome and returns a highly specific primer set

calculated using a precise physicochemical model. To shorten the response time,

we precomputed candidates for specific primers using a massively parallel

computer with 100 CPUs (SunFire 15 K) about 3 months in advance. This enables

PrimerStation to search and output qualified primers interactively. PrimerStation

can select highly specific primers suitable for multiplex PCR by seeking a wider

temperature range that minimizes the possibility of cross-reaction. It also

allows users to add heuristic rules to the primer design, e.g. the exclusion of

single nucleotide polymorphisms (SNPs) in primers, the avoidance of poly(A) and

CA-repeats in the PCR products, and the elimination of defective primers using

the secondary structure prediction. We performed several tests to verify the PCR

amplification of randomly selected primers for ChrX, and we confirmed that the

primers amplify specific PCR products perfectly.

DOI: 10.1093/nar/gkl297

PMCID: PMC1538814

PMID: 16845094 [Indexed for MEDLINE]

2884. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W656-9.

SSRPrimer and SSR Taxonomy Tree: Biome SSR discovery.

Jewell E(1), Robinson A, Savage D, Erwin T, Love CG, Lim GA, Li X, Batley J,

Spangenberg GC, Edwards D.

Author information:

(1)Department of Primary Industries, Plant Biotechnology Centre, Primary

Industries Research Victoria, Victorian AgriBiosciences Centre, 1 Park Drive,

Bundoora, Victoria 3083, Australia.

Simple sequence repeat (SSR) molecular genetic markers have become important

tools for a broad range of applications such as genome mapping and genetic

diversity studies. SSRs are readily identified within DNA sequence data and PCR

primers can be designed for their amplification. These PCR primers frequently

cross amplify within related species. We report a web-based tool, SSR Primer,

that integrates SPUTNIK, an SSR repeat finder, with Primer3, a primer design

program, within one pipeline. On submission of multiple FASTA formatted

sequences, the script screens each sequence for SSRs using SPUTNIK. Results are

then parsed to Primer3 for locus specific primer design. We have applied this

tool for the discovery of SSRs within the complete GenBank database, and have

designed PCR amplification primers for over 13 million SSRs. The SSR Taxonomy

Tree server provides web-based searching and browsing of species and taxa for the

visualisation and download of these SSR amplification primers. These tools are

available at http://bioinformatics.pbcbasc.latrobe.edu.au/ssrdiscovery.html.

DOI: 10.1093/nar/gkl083

PMCID: PMC1538772

PMID: 16845092 [Indexed for MEDLINE]

2885. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W651-5.

SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes.

Andreson R(1), Puurand T, Remm M.

Author information:

(1)Department of Bioinformatics, University of Tartu, Estonia.

SNPmasker is a comprehensive web interface for masking large eukaryotic genomes.

The program is designed to mask SNPs from recent dbSNP database and to mask the

repeats with two alternative programs. In addition to the SNP masking, we also

offer population-specific substitution of SNP alleles in genomic sequence

according to SNP frequencies in HapMap Phase II data. The input to SNPmasker can

be defined in chromosomal coordinates or inserted as a sequence. The sequences

masked by our web server are most useful as a preliminary step for different

primer and probe design tasks. The service is available at

http://bioinfo.ebc.ee/snpmasker/ and is free for all users.

DOI: 10.1093/nar/gkl125

PMCID: PMC1538889

PMID: 16845091 [Indexed for MEDLINE]

2886. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W635-41.

FASTSNP: an always up-to-date and extendable service for SNP function analysis

and prioritization.

Yuan HY(1), Chiou JJ, Tseng WH, Liu CH, Liu CK, Lin YJ, Wang HH, Yao A, Chen YT,

Hsu CN.

Author information:

(1)Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan.

Single nucleotide polymorphism (SNP) prioritization based on the phenotypic risk

is essential for association studies. Assessment of the risk requires access to a

variety of heterogeneous biological databases and analytical tools. FASTSNP

(function analysis and selection tool for single nucleotide polymorphisms) is a

web server that allows users to efficiently identify and prioritize high-risk

SNPs according to their phenotypic risks and putative functional effects. A

unique feature of FASTSNP is that the functional effect information used for SNP

prioritization is always up-to-date, because FASTSNP extracts the information

from 11 external web servers at query time using a team of web wrapper agents.

Moreover, FASTSNP is extendable by simply deploying more Web wrapper agents. To

validate the results of our prioritization, we analyzed 1569 SNPs from the

SNP500Cancer database. The results show that SNPs with a high predicted risk

exhibit low allele frequencies for the minor alleles, consistent with a

well-known finding that a strong selective pressure exists for functional

polymorphisms. We have been using FASTSNP for 2 years and FASTSNP enables us to

discover a novel promoter polymorphism. FASTSNP is available at

http://fastsnp.ibms.sinica.edu.tw.

DOI: 10.1093/nar/gkl236

PMCID: PMC1538865

PMID: 16845089 [Indexed for MEDLINE]

2887. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W63-9.

PREDITOR: a web server for predicting protein torsion angle restraints.

Berjanskii MV(1), Neal S, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB, Canada.

Every year between 500 and 1000 peptide and protein structures are determined by

NMR and deposited into the Protein Data Bank. However, the process of NMR

structure determination continues to be a manually intensive and time-consuming

task. One of the most tedious and error-prone aspects of this process involves

the determination of torsion angle restraints including phi, psi, omega and chi

angles. Most methods require many days of additional experiments, painstaking

measurements or complex calculations. Here we wish to describe a web server,

called PREDITOR, which greatly accelerates and simplifies this task. PREDITOR

accepts sequence and/or chemical shift data as input and generates torsion angle

predictions (with predicted errors) for phi, psi, omega and chi-1 angles.

PREDITOR combines sequence alignment methods with advanced chemical shift

analysis techniques to generate its torsion angle predictions. The method is fast

(<40 s per protein) and accurate, with 88% of phi/psi predictions being within 30

degrees of the correct values, 84% of chi-1 predictions being correct and 99.97%

of omega angles being correct. PREDITOR is 35 times faster and up to 20% more

accurate than any existing method. PREDITOR also provides accurate assessments of

the torsion angle errors so that the torsion angle constraints can be readily fed

into standard structure refinement programs, such as CNS, XPLOR, AMBER and CYANA.

Other unique features to PREDITOR include dihedral angle prediction via PDB

structure mapping, automated chemical shift re-referencing (to improve accuracy),

prediction of proline cis/trans states and a simple user interface. The PREDITOR

website is located at: http://wishart.biology.ualberta.ca/preditor.

DOI: 10.1093/nar/gkl341

PMCID: PMC1538894

PMID: 16845087 [Indexed for MEDLINE]

2888. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W617-20.

FootPrinter3: phylogenetic footprinting in partially alignable sequences.

Fang F(1), Blanchette M.

Author information:

(1)McGill Centre for Bioinformatics, 3775 University St, Montréal, Québec,

Canada, H3A 2B4.

FootPrinter3 is a web server for predicting transcription factor binding sites by

using phylogenetic footprinting. Until now, phylogenetic footprinting approaches

have been based either on multiple alignment analysis (e.g. PhyloVista,

PhastCons), or on motif-discovery algorithms (e.g. FootPrinter2). FootPrinter3

integrates these two approaches, making use of local multiple sequence alignment

blocks when those are available and reliable, but also allowing finding motifs in

unalignable regions. The result is a set of predictions that joins the advantages

of alignment-based methods (good specificity) to those of motif-based methods

(good sensitivity, even in the presence of highly diverged species). FootPrinter3

is thus a tool of choice to exploit the wealth of vertebrate genomes being

sequenced, as it allows taking full advantage of the sequences of highly diverged

species (e.g. chicken, zebrafish), as well as those of more closely related

species (e.g. mammals). The FootPrinter3 web server is available at:

http://www.mcb.mcgill.ca/~blanchem/FootPrinter3.

DOI: 10.1093/nar/gkl123

PMCID: PMC1538810

PMID: 16845084 [Indexed for MEDLINE]

2889. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W613-6.

AliWABA: alignment on the web through an A-Bruijn approach.

Jones NC(1), Zhi D, Raphael BJ.

Author information:

(1)Department of Computer Science and Engineering, University of California San

Diego, La Jolla, CA, USA. ncjones@cs.uscd.edu

Multiple sequence alignment programs are an invaluable tool in computational

biology. A-Bruijn Alignment (ABA) is a method for multiple sequence alignment

that represents an alignment as a directed graph and has proved useful in

aligning nucleotide and amino acid sequences that are composed of repeated and

shuffled subsequences. AliWABA is a web server that provides tools to generate

alignments with ABA, visualize the resulting ABA graphs and extract subsequences

from ABA graphs. AliWABA greatly simplifies the problem of analyzing multiple

sequences for local similarities that may be reordered, as is common with the

domain architectures of proteins. To facilitate the analysis of protein domains,

AliWABA provides direct querying of the Conserved Domain Database.AVAILABILITY:

http://aba.nbcr.net/

DOI: 10.1093/nar/gkl288

PMCID: PMC1538870

PMID: 16845083 [Indexed for MEDLINE]

2890. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W609-12.

PAL2NAL: robust conversion of protein sequence alignments into the corresponding

codon alignments.

Suyama M(1), Torrents D, Bork P.

Author information:

(1)European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg,

Germany.

PAL2NAL is a web server that constructs a multiple codon alignment from the

corresponding aligned protein sequences. Such codon alignments can be used to

evaluate the type and rate of nucleotide substitutions in coding DNA for a wide

range of evolutionary analyses, such as the identification of levels of selective

constraint acting on genes, or to perform DNA-based phylogenetic studies. The

server takes a protein sequence alignment and the corresponding DNA sequences as

input. In contrast to other existing applications, this server is able to

construct codon alignments even if the input DNA sequence has mismatches with the

input protein sequence, or contains untranslated regions and polyA tails. The

server can also deal with frame shifts and inframe stop codons in the input

models, and is thus suitable for the analysis of pseudogenes. Another distinct

feature is that the user can specify a subregion of the input alignment in order

to specifically analyze functional domains or exons of interest. The PAL2NAL

server is available at http://www.bork.embl.de/pal2nal.

DOI: 10.1093/nar/gkl315

PMCID: PMC1538804

PMID: 16845082 [Indexed for MEDLINE]

2891. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W604-8.

Expresso: automatic incorporation of structural information in multiple sequence

alignments using 3D-Coffee.

Armougom F(1), Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V,

Notredame C.

Author information:

(1)Laboratoire Information Génomique et Structurale, CNRS UPR2589, Institute for

Structural Biology and Microbiology (IBSM), Parc Scientifique de Luminy, 163

Avenue de Luminy, FR- 13288, Marseille cedex 09, France.

Expresso is a multiple sequence alignment server that aligns sequences using

structural information. The user only needs to provide sequences. The server runs

BLAST to identify close homologues of the sequences within the PDB database.

These PDB structures are used as templates to guide the alignment of the original

sequences using structure-based sequence alignment methods like SAP or Fugue. The

final result is a multiple sequence alignment of the original sequences based on

the structural information of the templates. An advanced mode makes it possible

to either upload private structures or specify which PDB templates should be used

to model each sequence. Providing the suitable structural information is

available, Expresso delivers sequence alignments with accuracy comparable with

structure-based alignments. The server is available on http://www.tcoffee.org/.

DOI: 10.1093/nar/gkl092

PMCID: PMC1538866

PMID: 16845081 [Indexed for MEDLINE]

2892. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W600-3.

PROTOGENE: turning amino acid alignments into bona fide CDS nucleotide

alignments.

Moretti S(1), Reinier F, Poirot O, Armougom F, Audic S, Keduas V, Notredame C.

Author information:

(1)Information Génomique et Structurale, CNRS UPR2589, Institute for Structural

Biology and Microbiology (IBSM), Parc Scientifique de Luminy, 163 Avenue de

Luminy, FR 13288, Marseille cedex 09, France.

We describe Protogene, a server that can turn a protein multiple sequence

alignment into the equivalent alignment of the original gene coding DNA.

Protogene relies on a pipeline where every initial protein sequence is BLASTed

against RefSeq or NR. The annotation associated with potential matches is used to

identify the gene sequence. This gene sequence is then aligned with the query

protein using Exonerate in order to extract a coding nucleotide sequence matching

the original protein. Protogene can handle protein fragments and will return

every CDS coding for a given protein, even if they occur in different genomes.

Protogene is available from http://www.tcoffee.org/.

DOI: 10.1093/nar/gkl170

PMCID: PMC1538918

PMID: 16845080 [Indexed for MEDLINE]

2893. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W6-9.

BLAST: improvements for better sequence analysis.

Ye J(1), McGinnis S, Madden TL.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD

20894, USA.

Basic local alignment search tool (BLAST) is a sequence similarity search

program. The National Center for Biotechnology Information (NCBI) maintains a

BLAST server with a home page at http://www.ncbi.nlm.nih.gov/BLAST/. We report

here on recent enhancements to the results produced by the BLAST server at the

NCBI. These include features to highlight mismatches between similar sequences,

show where the query was masked for low-complexity sequence, and integrate

information about the database sequences from the NCBI Entrez system into the

BLAST display. Changes to how the database sequences are fetched have also

improved the speed of the report generator.

DOI: 10.1093/nar/gkl164

PMCID: PMC1538791

PMID: 16845079 [Indexed for MEDLINE]

2894. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W578-83.

PromAn: an integrated knowledge-based web server dedicated to promoter analysis.

Lardenois A(1), Chalmel F, Bianchetti L, Sahel JA, Léveillard T, Poch O.

Author information:

(1)Laboratoire de Biologie et Génomique Structurales, Institut de Génétique et de

Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP 163, 67404 Illkirch

Cedex, France.

PromAn is a modular web-based tool dedicated to promoter analysis that integrates

distinct complementary databases, methods and programs. PromAn provides automatic

analysis of a genomic region with minimal prior knowledge of the genomic

sequence. Prediction programs and experimental databases are combined to locate

the transcription start site (TSS) and the promoter region within a large genomic

input sequence. Transcription factor binding sites (TFBSs) can be predicted using

several public databases and user-defined motifs. Also, a phylogenetic

footprinting strategy, combining multiple alignment of large genomic sequences

and assignment of various scores reflecting the evolutionary selection pressure,

allows for evaluation and ranking of TFBS predictions. PromAn results can be

displayed in an interactive graphical user interface, PromAnGUI. It integrates

all of this information to highlight active promoter regions, to identify among

the huge number of TFBS predictions those which are the most likely to be

potentially functional and to facilitate user refined analysis. Such an

integrative approach is essential in the face of a growing number of tools

dedicated to promoter analysis in order to propose hypotheses to direct further

experimental validations. PromAn is publicly available at

http://bips.u-strasbg.fr/PromAn.

DOI: 10.1093/nar/gkl193

PMCID: PMC1538850

PMID: 16845074 [Indexed for MEDLINE]

2895. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W571-7.

CRSD: a comprehensive web server for composite regulatory signature discovery.

Liu CC(1), Lin CC, Chen WS, Chen HY, Chang PC, Chen JJ, Yang PC.

Author information:

(1)Department of Computer Science, National Chung-Hsing University, Taichung,

Taiwan, ROC.

Transcription factors (TFs) and microRNAs play important roles in the regulation

of human gene expression, and the study of their combinatory regulations of gene

expression is a new research field. We constructed a comprehensive web server,

the composite regulatory signature database (CRSD), that can be applied in

investigating complex regulatory behaviors involving gene expression signatures

(GESs), microRNA regulatory signatures (MRSs) and TF regulatory signatures

(TRSs). Six well-known and large-scale databases, including the human UniGene,

mature microRNAs, putative promoter, TRANSFAC, pathway and Gene Ontology (GO)

databases, were integrated to provide the comprehensive analysis in CRSD. Two new

genome-wide databases, of MRSs and TRSs, were also constructed and further

integrated into CRSD. To accomplish the microarray data analysis at one go,

several methods, including microarray data pretreatment, statistical and

clustering analysis, iterative enrichment analysis and motif discovery, were

closely integrated in the web server, which has not been the case in previous

studies. Our implementation showed that the published literature could

demonstrate the results of genome-wide enrichment analysis. We conclude that CRSD

is a powerful and useful bioinformatic web server and may provide new insights

into gene regulation networks. CRSD and the online tutorial are publicly

available at http://biochip.nchu.edu.tw/crsd1/.

DOI: 10.1093/nar/gkl279

PMCID: PMC1538777

PMID: 16845073 [Indexed for MEDLINE]

2896. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W57-62.

UMMS: constrained harmonic and anharmonic analyses of macromolecules based on

elastic network models.

Jang Y(1), Jeong JI, Kim MK.

Author information:

(1)Department of Mechanical and Industrial Engineering, University of

Massachusetts, Amherst, MA 01003, USA.

UMass Morph Server (UMMS) has been developed for the broad impact on the study of

molecular dynamics (MD). The elastic network model (ENM) of a given macromolecule

has been proven as a useful tool for analyzing thermal behaviors locally and

predicting folding pathways globally. UMMS utilizes coarse-grained ENMs at

various levels. These simplifications remarkably save computation time compared

with all-atom MD simulations so that one can bring down massive computational

problems from a supercomputer to a PC. To improve computational efficiency and

physical reality of ENMs, the symmetry-constrained, rigid-cluster, hybrid and

chemical-bond ENMs have been developed and implemented at UMMS. One can request

both harmonic normal mode analysis of a single macromolecule and anharmonic

pathway generation between two conformations of a same molecule using elastic

network interpolation at http://biomechanics.ecs.umass.edu/umms.html.

DOI: 10.1093/nar/gkl039

PMCID: PMC1538792

PMID: 16845072 [Indexed for MEDLINE]

2897. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W566-70.

MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated

or homologous genes.

Pavesi G(1), Mereghetti P, Zambelli F, Stefani M, Mauri G, Pesole G.

Author information:

(1)Dipartimento di Scienze Biomolecolari e Biotecnologie, University of Milano,

Milano, Italy.

Understanding the complex mechanisms regulating gene expression at the

transcriptional and post-transcriptional levels is one of the greatest challenges

of the post-genomic era. The MoD (MOtif Discovery) Tools web server comprises a

set of tools for the discovery of novel conserved sequence and structure motifs

in nucleotide sequences, motifs that in turn are good candidates for regulatory

activity. The server includes the following programs: Weeder, for the discovery

of conserved transcription factor binding sites (TFBSs) in nucleotide sequences

from co-regulated genes; WeederH, for the discovery of conserved TFBSs and distal

regulatory modules in sequences from homologous genes; RNAProfile, for the

discovery of conserved secondary structure motifs in unaligned RNA sequences

whose secondary structure is not known. In this way, a given gene can be compared

with other co-regulated genes or with its homologs, or its mRNA can be analyzed

for conserved motifs regulating its post-transcriptional fate. The web server

thus provides researchers with different strategies and methods to investigate

the regulation of gene expression, at both the transcriptional and

post-transcriptional levels. Available at http://www.pesolelab.it/modtools/ and

http://www.beacon.unimi.it/modtools/.

DOI: 10.1093/nar/gkl285

PMCID: PMC1538899

PMID: 16845071 [Indexed for MEDLINE]

2898. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W555-9.

Stubb: a program for discovery and analysis of cis-regulatory modules.

Sinha S(1), Liang Y, Siggia E.

Author information:

(1)Department of Computer Science, University of Illinois, Urbana-Champaign, NY,

USA. sinhas@cs.uiuc.edu

Given the DNA-binding specificities (motifs) of one or more transcription

factors, an important bioinformatics problem is to discover significant clusters

of binding sites for the transcription factors(s). Such clusters often correspond

to cis-regulatory modules mediating regulation of an adjacent gene. In earlier

work, we developed the Stubb program that uses a probabilistic model and a

maximum likelihood approach to efficiently detect cis-regulatory modules over

genomic scales. It may optionally exploit a second related genome to improve

module prediction accuracy. We describe here the use of a web-based interface for

the Stubb program. The interface is equipped with a special post-processing step

for in-depth analysis of specific modules, in order to reveal individual binding

sites predicted in the module. The web server may be accessed at the URL

http://stubb.rockefeller.edu/.

DOI: 10.1093/nar/gkl224

PMCID: PMC1538799

PMID: 16845069 [Indexed for MEDLINE]

2899. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W551-4.

CEAS: cis-regulatory element annotation system.

Ji X(1), Li W, Song J, Wei L, Liu XS.

Author information:

(1)Center for Bioinformatics, National Laboratory of Protein Engineering and

Plant Genetic Engineering, College of Life Sciences, Peking University Beijing,

People's Republic China 100871.

The recent availability of high-density human genome tiling arrays enables

biologists to conduct ChIP-chip experiments to locate the in vivo-binding sites

of transcription factors in the human genome and explore the regulatory

mechanisms. Once genomic regions enriched by transcription factor ChIP-chip are

located, genome-scale downstream analyses are crucial but difficult for

biologists without strong bioinformatics support. We designed and implemented the

first web server to streamline the ChIP-chip downstream analyses. Given

genome-scale ChIP regions, the cis-regulatory element annotation system (CEAS)

retrieves repeat-masked genomic sequences, calculates GC content, plots

evolutionary conservation, maps nearby genes and identifies enriched

transcription factor-binding motifs. Biologists can utilize CEAS to retrieve

useful information for ChIP-chip validation, assemble important knowledge to

include in their publication and generate novel hypotheses (e.g. transcription

factor cooperative partner) for further study. CEAS helps the adoption of

ChIP-chip in mammalian systems and provides insights towards a more comprehensive

understanding of transcriptional regulatory mechanisms. The URL of the server is

http://ceas.cbi.pku.edu.cn.

DOI: 10.1093/nar/gkl322

PMCID: PMC1538818

PMID: 16845068 [Indexed for MEDLINE]

2900. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W541-5.

Composite Module Analyst: identification of transcription factor binding site

combinations using genetic algorithm.

Waleev T(1), Shtokalo D, Konovalova T, Voss N, Cheremushkin E, Stegmaier P,

Kel-Margoulis O, Wingender E, Kel A.

Author information:

(1)A.P. Ershov's Institute of Informatics Systems 6, Lavrentiev avenue, 630090

Novosibirsk, Russia.

Composite Module Analyst (CMA) is a novel software tool aiming to identify

promoter-enhancer models based on the composition of transcription factor (TF)

binding sites and their pairs. CMA is closely interconnected with the TRANSFAC

database. In particular, CMA uses the positional weight matrix (PWM) library

collected in TRANSFAC and therefore provides the possibility to search for a

large variety of different TF binding sites. We model the structure of the long

gene regulatory regions by a Boolean function that joins several local modules,

each consisting of co-localized TF binding sites. Having as an input a set of

co-regulated genes, CMA builds the promoter model and optimizes the parameters of

the model automatically by applying a genetic-regression algorithm. We use a

multicomponent fitness function of the algorithm which includes several

statistical criteria in a weighted linear function. We show examples of

successful application of CMA to a microarray data on transcription profiling of

TNF-alpha stimulated primary human endothelial cells. The CMA web server is

freely accessible at http://www.gene-regulation.com/pub/programs/cma/CMA.html. An

advanced version of CMA is also a part of the commercial system ExPlaintrade mark

(www.biobase.de) designed for causal analysis of gene expression data.

DOI: 10.1093/nar/gkl342

PMCID: PMC1538785

PMID: 16845066 [Indexed for MEDLINE]

2901. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W529-33.

VOMBAT: prediction of transcription factor binding sites using variable order

Bayesian trees.

Grau J(1), Ben-Gal I, Posch S, Grosse I.

Author information:

(1)Institute of Computer Science, University Halle, 06099 Halle, Saale, Germany.

Variable order Markov models and variable order Bayesian trees have been proposed

for the recognition of transcription factor binding sites, and it could be

demonstrated that they outperform traditional models, such as position weight

matrices, Markov models and Bayesian trees. We develop a web server for the

recognition of DNA binding sites based on variable order Markov models and

variable order Bayesian trees offering the following functionality: (i) given

datasets with annotated binding sites and genomic background sequences, variable

order Markov models and variable order Bayesian trees can be trained; (ii) given

a set of trained models, putative DNA binding sites can be predicted in a given

set of genomic sequences and (iii) given a dataset with annotated binding sites

and a dataset with genomic background sequences, cross-validation experiments for

different model combinations with different parameter settings can be performed.

Several of the offered services are computationally demanding, such as

genome-wide predictions of DNA binding sites in mammalian genomes or sets of

10(4)-fold cross-validation experiments for different model combinations based on

problem-specific data sets. In order to execute these jobs, and in order to serve

multiple users at the same time, the web server is attached to a Linux cluster

with 150 processors. VOMBAT is available at

http://pdw-24.ipk-gatersleben.de:8080/VOMBAT/.

DOI: 10.1093/nar/gkl212

PMCID: PMC1538886

PMID: 16845064 [Indexed for MEDLINE]

2902. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W524-8.

TFBScluster web server for the identification of mammalian composite regulatory

elements.

Donaldson IJ(1), Göttgens B.

Author information:

(1)Department of Haematology, Cambridge Institute for Medical Research,

University of Cambridge, Hills Road, Cambridge, CB2 2XY, UK.

Identification of transcriptional regulatory elements represents a critical step

in our ability to reconstruct transcriptional regulatory networks from gene

expression profiling datasets. To facilitate computational identification of

candidate gene regulatory elements from whole genome sequences, we have developed

the TFBScluster web server that integrates several tools for the genome-wide

identification and subsequent characterization of transcription factor binding

site clusters that are conserved in multiple mammalian species. Either the human

or mouse genomes can be used as the reference sequence with direct links from the

search results to the ENSEMBL and UCSC genome browsers. Moreover, TFBScluster

provides seamless integration of transcription factor binding site searches with

genome annotation and gene expression profiling data, to allow prioritising

computational predictions for subsequent experimental validation. TFBScluster is

publicly available at http://hscl.cimr.cam.ac.uk/TFBScluster\_genome\_portal.html.

DOI: 10.1093/nar/gkl041

PMCID: PMC1538905

PMID: 16845063 [Indexed for MEDLINE]

2903. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W52-6.

NOMAD-Ref: visualization, deformation and refinement of macromolecular structures

based on all-atom normal mode analysis.

Lindahl E(1), Azuara C, Koehl P, Delarue M.

Author information:

(1)Unite de Dynamique Structurale des Macromolécules, URA 2185 du C.N.R.S.,

Institut Pasteur, 75015 Paris, France.

Normal mode analysis (NMA) is an efficient way to study collective motions in

biomolecules that bypasses the computational costs and many limitations

associated with full dynamics simulations. The NOMAD-Ref web server presented

here provides tools for online calculation of the normal modes of large molecules

(up to 100,000 atoms) maintaining a full all-atom representation of their

structures, as well as access to a number of programs that utilize these

collective motions for deformation and refinement of biomolecular structures.

Applications include the generation of sets of decoys with correct

stereochemistry but arbitrary large amplitude movements, the quantification of

the overlap between alternative conformations of a molecule, refinement of

structures against experimental data, such as X-ray diffraction structure factors

or Cryo-EM maps and optimization of docked complexes by modeling receptor/ligand

flexibility through normal mode motions. The server can be accessed at the URL

http://lorentz.immstr.pasteur.fr/nomad-ref.php.

DOI: 10.1093/nar/gkl082

PMCID: PMC1538881

PMID: 16845062 [Indexed for MEDLINE]

2904. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W504-9.

Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene

expression analysis.

Manfield IW(1), Jen CH, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM,

Westhead DR.

Author information:

(1)Centre for Plant Sciences, University of Leeds, West Yorkshire, LS2 9JT, UK.

The Arabidopsis Co-expression Tool, ACT, ranks the genes across a large

microarray dataset according to how closely their expression follows the

expression of a query gene. A database stores pre-calculated co-expression

results for approximately 21,800 genes based on data from over 300 arrays. These

results can be corroborated by calculation of co-expression results for

user-defined sub-sets of arrays or experiments from the NASC/GARNet array

dataset. Clique Finder (CF) identifies groups of genes which are consistently

co-expressed with each other across a user-defined co-expression list. The

parameters can be altered easily to adjust cluster size and the output examined

for optimal inclusion of genes with known biological roles. Alternatively, a

Scatter Plot tool displays the correlation coefficients for all genes against two

user-selected queries on a scatter plot which can be useful for visual

identification of clusters of genes with similar r-values. User-input groups of

genes can be highlighted on the scatter plots. Inclusion of genes with known

biology in sets of genes identified using CF and Scatter Plot tools allows

inferences to be made about the roles of the other genes in the set and both

tools can therefore be used to generate short lists of genes for further

characterization. ACT is freely available at www.Arabidopsis.leeds.ac.uk/ACT.

DOI: 10.1093/nar/gkl204

PMCID: PMC1538833

PMID: 16845059 [Indexed for MEDLINE]

2905. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W492-7.

GEPS: the Gene Expression Pattern Scanner.

Wang YP(1), Liang L, Han BC, Quan Y, Wang X, Tao T, Ji ZL.

Author information:

(1)Key Laboratory for Cell Biology and Tumor Cell Engineering, the Ministry of

Education of China, School of Life Sciences, Xiamen University, Xiamen 361005,

Fujian, People's Republic of China.

Gene Expression Pattern Scanner (GEPS) is a web-based server to provide

interactive pattern analysis of user-submitted microarray data for facilitating

their further interpretation. Putative gene expression patterns such as

correlated expression, similar expression and specific expression are determined

globally and systematically using geometric comparison and correlation analysis

methods. These patterns can be visualized via linear plot with quantitative

measures. User-defined threshold value is allowed to customize the format of the

pattern search results. For better understanding of gene expression, patterns

derived from 329,205 non-redundant gene expression records from the GNF SymAltas

and the Gene Expression Omnibus are also provided. These profiles cover 24,277

human genes in 79 tissues, 32,905 mouse genes in 61 tissues and 4201 rat genes in

44 tissues. GEPS is available at http://bioinf.xmu.edu.cn/software/geps/geps.php.

DOI: 10.1093/nar/gkl067

PMCID: PMC1538815

PMID: 16845057 [Indexed for MEDLINE]

2906. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W482-5.

BTW: a web server for Boltzmann time warping of gene expression time series.

Ferrè F(1), Clote P.

Author information:

(1)Department of Biology, Boston College, Chestnut Hill, MA, USA.

Dynamic time warping (DTW) is a well-known quadratic time algorithm to determine

the smallest distance and optimal alignment between two numerical sequences,

possibly of different length. Originally developed for speech recognition, this

method has been used in data mining, medicine and bioinformatics. For gene

expression time series data, time warping distance is arguably a more flexible

tool to determine genes having similar temporal expression, hence possibly

related biological function, than either Euclidean distance or correlation

coefficient--especially since time warping accommodates sequences of different

length. The BTW web server allows a user to upload two tab-separated text files

A,B of gene expression data, each possibly having a different number of time

intervals of different durations. BTW then computes time warping distance between

each gene of A with each gene of B, using a recently developed symmetric

algorithm which additionally computes the Boltzmann partition function and

outputs Boltzmann pair probabilities. The Boltzmann pair probabilities, not

available with any other existent software, suggest possible biological

significance of certain positions in an optimal time warping

alignment.AVAILABILITY: http://bioinformatics.bc.edu/clotelab/BTW/.

DOI: 10.1093/nar/gkl162

PMCID: PMC1538860

PMID: 16845055 [Indexed for MEDLINE]

2907. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W48-51.

pKD: re-designing protein pKa values.

Tynan-Connolly BM(1), Nielsen JE.

Author information:

(1)School of Biomolecular and Biomedical Science, Centre for Synthesis and

Chemical Biology, UCD Conway Institute, University College Dublin, Belfield,

Dublin 4, Ireland.

The pK(a) values in proteins govern the pH-dependence of protein stability and

enzymatic activity. A large number of mutagenesis experiments have been carried

out in the last three decades to re-engineer the pH-activity and pH-stability

profile of enzymes and proteins. We have developed the pKD webserver

(http://polymerase.ucd.ie/pKa\_Design), which predicts sets of point mutations

that will change the pK(a) values of a set of target residues in a given

direction, thus allowing for targeted re-design of the pH-dependent

characteristics of proteins. The server provides the user with an interactive

experience for re-designing pK(a) values by pre-calculating DeltapK(a) values

from all feasible point mutations. Design solutions are found in less than 10 min

for a typical design job for a medium-sized protein. Mutant DeltapK(a) values

calculated by the pKD web server are in close agreement with those produced by

comparing results from full-fledged pK(a) calculation methods.

DOI: 10.1093/nar/gkl192

PMCID: PMC1538816

PMID: 16845054 [Indexed for MEDLINE]

2908. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W466-71.

BiologicalNetworks: visualization and analysis tool for systems biology.

Baitaluk M(1), Sedova M, Ray A, Gupta A.

Author information:

(1)San Diego Supercomputer Center, University of California San Diego, 9500

Gilman Drive, La Jolla, CA 92093, USA. baitaluk@sdsc.edu

Systems level investigation of genomic scale information requires the development

of truly integrated databases dealing with heterogeneous data, which can be

queried for simple properties of genes or other database objects as well as for

complex network level properties, for the analysis and modelling of complex

biological processes. Towards that goal, we recently constructed PathSys, a data

integration platform for systems biology, which provides dynamic integration over

a diverse set of databases [Baitaluk et al. (2006) BMC Bioinformatics 7, 55].

Here we describe a server, BiologicalNetworks, which provides visualization,

analysis services and an information management framework over PathSys. The

server allows easy retrieval, construction and visualization of complex

biological networks, including genome-scale integrated networks of

protein-protein, protein-DNA and genetic interactions. Most importantly,

BiologicalNetworks addresses the need for systematic presentation and analysis of

high-throughput expression data by mapping and analysis of expression profiles of

genes or proteins simultaneously on to regulatory, metabolic and cellular

networks. BiologicalNetworks Server is available at

http://brak.sdsc.edu/pub/BiologicalNetworks.

DOI: 10.1093/nar/gkl308

PMCID: PMC1538788

PMID: 16845051 [Indexed for MEDLINE]

2909. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W463-5.

jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1.

Zhang M(1), Schultz AK, Calef C, Kuiken C, Leitner T, Korber B, Morgenstern B,

Stanke M.

Author information:

(1)Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik,

Goldschmidtstrasse 1, 37077 Göttingen, Germany.

Detecting recombinations in the genome sequence of human immunodeficiency virus

(HIV-1) is crucial for epidemiological studies and for vaccine development.

Herein, we present a web server for subtyping and localization of phylogenetic

breakpoints in HIV-1. Our software is based on a jumping profile Hidden Markov

Model (jpHMM), a probabilistic generalization of the jumping-alignment approach

proposed by Spang et al. The input data for our server is a partial or complete

genome sequence from HIV-1; our tool assigns regions of the input sequence to

known subtypes of HIV-1 and predicts phylogenetic breakpoints. jpHMM is available

online at http://jphmm.gobics.de/.

DOI: 10.1093/nar/gkl255

PMCID: PMC1538796

PMID: 16845050 [Indexed for MEDLINE]

2910. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W459-62.

EGassembler: online bioinformatics service for large-scale processing, clustering

and assembling ESTs and genomic DNA fragments.

Masoudi-Nejad A(1), Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa

M, Endo T, Goto S.

Author information:

(1)Laboratory of Bioknowledge Systems, Bioinformatics Center, Institute for

Chemical Research, Kyoto University, Gokasho Uji, Kyoto 611-0011, Japan.

amasoudi@kais.kyoto-u.ac.jp

Expressed sequence tag (EST) sequencing has proven to be an economically feasible

alternative for gene discovery in species lacking a draft genome sequence.

Ongoing large-scale EST sequencing projects feel the need for bioinformatics

tools to facilitate uniform EST handling. This brings about a renewed importance

for a universal tool for processing and functional annotation of large sets of

ESTs. EGassembler (http://egassembler.hgc.jp/) is a web server, which provides an

automated as well as a user-customized analysis tool for cleaning, repeat

masking, vector trimming, organelle masking, clustering and assembling of ESTs

and genomic fragments. The web server is publicly available and provides the

community a unique all-in-one online application web service for large-scale ESTs

and genomic DNA clustering and assembling. Running on a Sun Fire 15K

supercomputer, a significantly large volume of data can be processed in a short

period of time. The results can be used to functionally annotate genes, to

facilitate splice alignment analysis, to link the transcripts to genetic and

physical maps, design microarray chips, to perform transcriptome analysis and to

map to KEGG metabolic pathways. The service provides an excellent bioinformatics

tool to research groups in wet-lab as well as an all-in-one-tool for sequence

handling to bioinformatics researchers.

DOI: 10.1093/nar/gkl066

PMCID: PMC1538775

PMID: 16845049 [Indexed for MEDLINE]

2911. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W455-8.

ProMiR II: a web server for the probabilistic prediction of clustered,

nonclustered, conserved and nonconserved microRNAs.

Nam JW(1), Kim J, Kim SK, Zhang BT.

Author information:

(1)Graduate Program in Bioinformatics, Seoul National University, Seoul 151-744,

Korea.

ProMiR is a web-based service for the prediction of potential microRNAs (miRNAs)

in a query sequence of 60-150 nt, using a probabilistic colearning model.

Identification of miRNAs requires a computational method to predict clustered and

nonclustered, conserved and nonconserved miRNAs in various species. Here we

present an improved version of ProMiR for identifying new clusters near known or

unknown miRNAs. This new version, ProMiR II, integrates additional evidence, such

as free energy data, G/C ratio, conservation score and entropy of candidate

sequences, for more controllable prediction of miRNAs in mouse and human genomes.

It also provides a wider range of services, e.g. the prediction of miRNA genes in

long nonrelated sequences such as viral genomes. Importantly, we have validated

this method using several case studies. All data used in ProMiR II are structured

in the MySQL database for efficient analysis. The ProMiR II web server is

available at http://cbit.snu.ac.kr/~ProMiR2/.

DOI: 10.1093/nar/gkl321

PMCID: PMC1538778

PMID: 16845048 [Indexed for MEDLINE]

2912. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W444-7.

ASGS: an alternative splicing graph web service.

Bollina D(1), Lee BT, Tan TW, Ranganathan S.

Author information:

(1)Department of Chemistry and Biomolecular Sciences and Biotechnology Research

Institute, Macquarie University, Sydney, NSW 2109, Australia.

Alternative transcript diversity manifests itself a prime cause of complexity in

higher eukaryotes. The Alternative Splicing Graph Server (ASGS) is a web service

facilitating the systematic study of alternatively spliced genes of higher

eukaryotes by generating splicing graphs for the compact visual representation of

transcript diversity from a single gene. Taking a set of transcripts in General

Feature Format as input, ASGS identifies distinct reference and variable exons,

generates a transcript splicing graph, an exon summary, splicing events

classification and a single line graph to facilitate experimental analysis. This

freely available web service can be accessed at http://asgs.biolinfo.org.

DOI: 10.1093/nar/gkl268

PMCID: PMC1538904

PMID: 16845045 [Indexed for MEDLINE]

2913. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W435-9.

AUGUSTUS: ab initio prediction of alternative transcripts.

Stanke M(1), Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B.

Author information:

(1)Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik,

Goldschmidtstrasse 1, 37077 Göttingen, Germany. mstanke@gwdg.de

AUGUSTUS is a software tool for gene prediction in eukaryotes based on a

Generalized Hidden Markov Model, a probabilistic model of a sequence and its gene

structure. Like most existing gene finders, the first version of AUGUSTUS

returned one transcript per predicted gene and ignored the phenomenon of

alternative splicing. Herein, we present a WWW server for an extended version of

AUGUSTUS that is able to predict multiple splice variants. To our knowledge, this

is the first ab initio gene finder that can predict multiple transcripts. In

addition, we offer a motif searching facility, where user-defined regular

expressions can be searched against putative proteins encoded by the predicted

genes. The AUGUSTUS web interface and the downloadable open-source stand-alone

program are freely available from http://augustus.gobics.de.

DOI: 10.1093/nar/gkl200

PMCID: PMC1538822

PMID: 16845043 [Indexed for MEDLINE]

2914. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W43-7.

PHEPS: web-based pH-dependent Protein Electrostatics Server.

Kantardjiev AA(1), Atanasov BP.

Author information:

(1)Biophysical Chemistry Group, Institute of Organic Chemistry, Bulgarian Academy

of Sciences, Sofia-1113, Bulgaria.

PHEPS (pH-dependent Protein Electrostatics Server) is a web service for fast

prediction and experiment planning support, as well as for correlation and

analysis of experimentally obtained results, reflecting charge-dependent

phenomena in globular proteins. Its implementation is based on long-term

experience (PHEI package) and the need to explain measured physicochemical

characteristics at the level of protein atomic structure. The approach is

semi-empirical and based on a mean field scheme for description and evaluation of

global and local pH-dependent electrostatic properties: protein proton binding;

ionic sites proton population; free energy electrostatic term; ionic groups

proton affinities (pK(a,i)) and their Coulomb interaction with whole charge

multipole; electrostatic potential of whole molecule at fixed pH and pH-dependent

local electrostatic potentials at user-defined set of points. The speed of

calculation is based on fast determination of distance-dependent pair

charge-charge interactions as empirical three exponential function that covers

charge-charge, charge-dipole and dipole-dipole contributions. After atomic

coordinates input, all standard parameters are used as defaults to facilitate

non-experienced users. Special attention was given to interactive addition of

non-polypeptide charges, extra ionizable groups with intrinsic pK(a)s or fixed

ions. The output information is given as plain-text, readable by 'RasMol',

'Origin' and the like. The PHEPS server is accessible at

http://pheps.orgchm.bas.bg/home.html.

DOI: 10.1093/nar/gkl165

PMCID: PMC1538834

PMID: 16845042 [Indexed for MEDLINE]

2915. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W429-34.

RegRNA: an integrated web server for identifying regulatory RNA motifs and

elements.

Huang HY(1), Chien CH, Jen KH, Huang HD.

Author information:

(1)Institute of Bioinformatics, National Chiao Tung University, Hsin-Chu 300,

Taiwan.

Numerous regulatory structural motifs have been identified as playing essential

roles in transcriptional and post-transcriptional regulation of gene expression.

RegRNA is an integrated web server for identifying the homologs of regulatory RNA

motifs and elements against an input mRNA sequence. Both sequence homologs and

structural homologs of regulatory RNA motifs can be recognized. The regulatory

RNA motifs supported in RegRNA are categorized into several classes: (i) motifs

in mRNA 5'-untranslated region (5'-UTR) and 3'-UTR; (ii) motifs involved in mRNA

splicing; (iii) motifs involved in transcriptional regulation; (iv) riboswitches;

(v) splicing donor/acceptor sites; (vi) inverted repeats; and (vii) miRNA target

sites. The experimentally validated regulatory RNA motifs are extracted from

literature survey and several regulatory RNA motif databases, such as UTRdb,

TRANSFAC, alternative splicing database (ASD) and miRBase. A variety of

computational programs are integrated for identifying the homologs of the

regulatory RNA motifs. An intuitive user interface is designed to facilitate the

comprehensive annotation of user-submitted mRNA sequences. The RegRNA web server

is now available at http://RegRNA.mbc.NCTU.edu.tw/.

DOI: 10.1093/nar/gkl333

PMCID: PMC1538840

PMID: 16845041 [Indexed for MEDLINE]

2916. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W423-8.

RNAMST: efficient and flexible approach for identifying RNA structural homologs.

Chang TH(1), Huang HD, Chuang TN, Shien DM, Horng JT.

Author information:

(1)Department of Computer Science and Information Engineering, National Central

University, Chung-Li 320, Taiwan.

RNA molecules fold into characteristic secondary structures for their diverse

functional activities such as post-translational regulation of gene expression.

Searching homologs of a pre-defined RNA structural motif, which may be a known

functional element or a putative RNA structural motif, can provide useful

information for deciphering RNA regulatory mechanisms. Since searching for the

RNA structural homologs among the numerous RNA sequences is extremely

time-consuming, this work develops a data preprocessing strategy to enhance the

search efficiency and presents RNAMST, which is an efficient and flexible web

server for rapidly identifying homologs of a pre-defined RNA structural motif

among numerous RNA sequences. Intuitive user interface are provided on the web

server to facilitate the predictive analysis. By comparing the proposed web

server to other tools developed previously, RNAMST performs remarkably more

efficiently and provides more effective and flexible functions. RNAMST is now

available on the web at http://bioinfo.csie.ncu.edu.tw/~rnamst/.

DOI: 10.1093/nar/gkl231

PMCID: PMC1538813

PMID: 16845040 [Indexed for MEDLINE]

2917. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W412-5.

The ARTS web server for aligning RNA tertiary structures.

Dror O(1), Nussinov R, Wolfson HJ.

Author information:

(1)School of Computer Science, Raymond and Beverly Sackler Faculty of Exact

Sciences, Tel Aviv University, Tel Aviv 69978, Israel. oranit@post.tau.ac.il

RNA molecules with common structural features may share similar functional

properties. Structural comparison of RNAs and detection of common substructures

is, thus, a highly important task. Nevertheless, the current available tools in

the RNA community provide only a partial solution, since they either work at the

2D level or are suitable for detecting predefined or local contiguous tertiary

motifs only. Here, we describe a web server built around ARTS, a method for

aligning tertiary structures of nucleic acids (both RNA and DNA). ARTS receives a

pair of 3D nucleic acid structures and searches for a priori unknown common

substructures. The search is truly 3D and irrespective of the order of the

nucleotides on the chain. The identified common substructures can be large global

folds with hundreds and even thousands of nucleotides as well as small local

motifs with at least two successive base pairs. The method is highly efficient

and has been used to conduct an all-against-all comparison of all the RNA

structures in the Protein Data Bank. The web server together with a software

package for download are freely accessible at http://bioinfo3d.cs.tau.ac.il/ARTS.

DOI: 10.1093/nar/gkl312

PMCID: PMC1538835

PMID: 16845038 [Indexed for MEDLINE]

2918. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W405-11.

CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments.

Bindewald E(1), Schneider TD, Shapiro BA.

Author information:

(1)Basic Research Program, SAIC-Frederick, NCI-Frederick, Frederick, MD 21702,

USA.

We present an online server that generates a 3D representation of properties of

user-submitted RNA or DNA alignments. The visualized properties are information

of single alignment columns, mutual information of two alignment positions as

well as the position-specific fraction of gaps. The nucleotide composition of

both single columns and column pairs is visualized with the help of color-coded

3D bars labeled with letters. The server generates both VRML and JVX output that

can be viewed with a VRML viewer or the JavaView applet, respectively. We show

that combining these different features of an alignment into one 3D

representation is helpful in identifying correlations between bases and potential

RNA and DNA base pairs. Significant known correlations between the tRNA 3'

anticodon cardinal nucleotide and the extended anticodon were observed, as were

correlations within the amino acid acceptor stem and between the cardinal

nucleotide and the acceptor stem. The online server can be accessed using the URL

http://correlogo.abcc.ncifcrf.gov.

DOI: 10.1093/nar/gkl269

PMCID: PMC1538790

PMID: 16845037 [Indexed for MEDLINE]

2919. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W394-9.

NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA

genes.

DeSantis TZ Jr(1), Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan

R, Andersen GL.

Author information:

(1)Lawrence Berkeley National Laboratory, Center for Environmental Biotechnology,

Berkeley, CA, USA.

Microbiologists conducting surveys of bacterial and archaeal diversity often

require comparative alignments of thousands of 16S rRNA genes collected from a

sample. The computational resources and bioinformatics expertise required to

construct such an alignment has inhibited high-throughput analysis. It was

hypothesized that an online tool could be developed to efficiently align

thousands of 16S rRNA genes via the NAST (Nearest Alignment Space Termination)

algorithm for creating multiple sequence alignments (MSA). The tool was

implemented with a web-interface at http://greengenes.lbl.gov/NAST. Each

user-submitted sequence is compared with Greengenes' 'Core Set', comprising

approximately 10,000 aligned non-chimeric sequences representative of the

currently recognized diversity among bacteria and archaea. User sequences are

oriented and paired with their closest match in the Core Set to serve as a

template for inserting gap characters. Non-16S data (sequence from vector or

surrounding genomic regions) are conveniently removed in the returned alignment.

From the resulting MSA, distance matrices can be calculated for diversity

estimates and organisms can be classified by taxonomy. The ability to align and

categorize large sequence sets using a simple interface has enabled researchers

with various experience levels to obtain bacterial and archaeal community

profiles.

DOI: 10.1093/nar/gkl244

PMCID: PMC1538769

PMID: 16845035 [Indexed for MEDLINE]

2920. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W389-93.

GenDecoder: genetic code prediction for metazoan mitochondria.

Abascal F(1), Zardoya R, Posada D.

Author information:

(1)Departamento de Bioquímica, Genética, e Inmunología, Universidad de Vigo,

36310 Vigo, Spain. fabascal@uvigo.es

Although the majority of the organisms use the same genetic code to translate

DNA, several variants have been described in a wide range of organisms, both in

nuclear and organellar systems, many of them corresponding to metazoan

mitochondria. These variants are usually found by comparative sequence analyses,

either conducted manually or with the computer. Basically, when a particular

codon in a query-species is linked to positions for which a specific amino acid

is consistently found in other species, then that particular codon is expected to

translate as that specific amino acid. Importantly, and despite the simplicity of

this approach, there are no available tools to help predicting the genetic code

of an organism. We present here GenDecoder, a web server for the characterization

and prediction of mitochondrial genetic codes in animals. The analysis of

automatic predictions for 681 metazoans aimed us to study some properties of the

comparative method, in particular, the relationship among sequence conservation,

taxonomic sampling and reliability of assignments. Overall, the method is highly

precise (99%), although highly divergent organisms such as platyhelminths are

more problematic. The GenDecoder web server is freely available from

http://darwin.uvigo.es/software/gendecoder.html.

DOI: 10.1093/nar/gkl044

PMCID: PMC1538875

PMID: 16845034 [Indexed for MEDLINE]

2921. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W382-4.

SWAKK: a web server for detecting positive selection in proteins using a sliding

window substitution rate analysis.

Liang H(1), Zhou W, Landweber LF.

Author information:

(1)Department of Chemistry, Princeton University, Princeton, NJ 08544, USA.

We present a bioinformatic web server (SWAKK) for detecting amino acid sites or

regions of a protein under positive selection. It estimates the ratio of

non-synonymous to synonymous substitution rates (K(A)/K(S)) between a pair of

protein-coding DNA sequences, by sliding a 3D window, or sphere, across one

reference structure. The program displays the results on the 3D protein

structure. In addition, for comparison or when a reference structure is

unavailable, the server can also perform a sliding window analysis on the primary

sequence. The SWAKK web server is available at

http://oxytricha.princeton.edu/SWAKK/.

DOI: 10.1093/nar/gkl272

PMCID: PMC1538794

PMID: 16845032 [Indexed for MEDLINE]

2922. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W38-42.

PDB\_Hydro: incorporating dipolar solvents with variable density in the

Poisson-Boltzmann treatment of macromolecule electrostatics.

Azuara C(1), Lindahl E, Koehl P, Orland H, Delarue M.

Author information:

(1)Unité de Dynamique Structurale des Macromolécules, URA 2185 du C.N.R.S.,

Institut Pasteur, 75015 Paris, France.

We describe a new way to calculate the electrostatic properties of macromolecules

which eliminates the assumption of a constant dielectric value in the solvent

region, resulting in a Generalized Poisson-Boltzmann-Langevin equation (GPBLE).

We have implemented a web server (http://lorentz.immstr.pasteur.fr/pdb\_hydro.php)

that both numerically solves this equation and uses the resulting water density

profiles to place water molecules at preferred sites of hydration. Surface atoms

with high or low hydration preference can be easily displayed using a simple

PyMol script, allowing for the tentative prediction of the dimerization interface

in homodimeric proteins, or lipid binding regions in membrane proteins. The web

site includes options that permit mutations in the sequence as well as

reconstruction of missing side chain and/or main chain atoms. These tools are

accessible independently from the electrostatics calculation, and can be used for

other modeling purposes. We expect this web server to be useful to structural

biologists, as the knowledge of solvent density should prove useful to get better

fits at low resolution for X-ray diffraction data and to computational

biologists, for whom these profiles could improve the calculation of interaction

energies in water between ligands and receptors in docking simulations.

DOI: 10.1093/nar/gkl072

PMCID: PMC1538897

PMID: 16845031 [Indexed for MEDLINE]

2923. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W379-81.

JAFA: a protein function annotation meta-server.

Friedberg I(1), Harder T, Godzik A.

Author information:

(1)Burnham Institute for Medical Research, Program in Bioinformatics and Systems

Biology, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA.

idoerg@burnham.org

With the high number of sequences and structures streaming in from genomic

projects, there is a need for more powerful and sophisticated annotation tools.

Most problematic of the annotation efforts is predicting gene and protein

function. Over the past few years there has been considerable progress in

automated protein function prediction, using a diverse set of methods.

Nevertheless, no single method reports all the information possible, and

molecular biologists resort to 'shopping around' using different methods: a

cumbersome and time-consuming practice. Here we present the Joined Assembly of

Function Annotations, or JAFA server. JAFA queries several function prediction

servers with a protein sequence and assembles the returned predictions in a

legible, non-redundant format. In this manner, JAFA combines the predictions of

several servers to provide a comprehensive view of what are the predicted

functions of the proteins. JAFA also offers its own output, and the individual

programs' predictions for further processing. JAFA is available for use from

http://jafa.burnham.org.

DOI: 10.1093/nar/gkl045

PMCID: PMC1538919

PMID: 16845030 [Indexed for MEDLINE]

2924. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W374-8.

HHsenser: exhaustive transitive profile search using HMM-HMM comparison.

Söding J(1), Remmert M, Biegert A, Lupas AN.

Author information:

(1)Department of Protein Evolution, Max-Planck-Institute for Developmental

Biology, Spemannstrasse 35, 72076 Tübingen, Germany.

johannes.soeding@tuebingen.mpg.de

HHsenser is the first server to offer exhaustive intermediate profile searches,

which it combines with pairwise comparison of hidden Markov models. Starting from

a single protein sequence or a multiple alignment, it can iteratively explore

whole superfamilies, producing few or no false positives. The output is a

multiple alignment of all detected homologs. HHsenser's sensitivity should make

it a useful tool for evolutionary studies. It may also aid applications that rely

on diverse multiple sequence alignments as input, such as homology-based

structure and function prediction, or the determination of functional residues by

conservation scoring and functional subtyping.HHsenser can be accessed at

http://hhsenser.tuebingen.mpg.de/. It has also been integrated into our structure

and function prediction server HHpred (http://hhpred.tuebingen.mpg.de/) to

improve predictions for near-singleton sequences.

DOI: 10.1093/nar/gkl195

PMCID: PMC1538784

PMID: 16845029 [Indexed for MEDLINE]

2925. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W369-73.

MEME: discovering and analyzing DNA and protein sequence motifs.

Bailey TL(1), Williams N, Misleh C, Li WW.

Author information:

(1)Institute of Molecular Bioscience, The University of Queensland, St Lucia, QLD

4072, Australia. t.bailey@imb.uq.edu.au

MEME (Multiple EM for Motif Elicitation) is one of the most widely used tools for

searching for novel 'signals' in sets of biological sequences. Applications

include the discovery of new transcription factor binding sites and protein

domains. MEME works by searching for repeated, ungapped sequence patterns that

occur in the DNA or protein sequences provided by the user. Users can perform

MEME searches via the web server hosted by the National Biomedical Computation

Resource (http://meme.nbcr.net) and several mirror sites. Through the same web

server, users can also access the Motif Alignment and Search Tool to search

sequence databases for matches to motifs encoded in several popular formats. By

clicking on buttons in the MEME output, users can compare the motifs discovered

in their input sequences with databases of known motifs, search sequence

databases for matches to the motifs and display the motifs in various formats.

This article describes the freely accessible web server and its architecture, and

discusses ways to use MEME effectively to find new sequence patterns in

biological sequences and analyze their significance.

DOI: 10.1093/nar/gkl198

PMCID: PMC1538909

PMID: 16845028 [Indexed for MEDLINE]

2926. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W350-5.

DILIMOT: discovery of linear motifs in proteins.

Neduva V(1), Russell RB.

Author information:

(1)EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

Discovery of protein functional motifs is critical in modern biology. Small

segments of 3-10 residues play critical roles in protein interactions,

post-translational modifications and trafficking. DILIMOT (DIscovery of LInear

MOTifs) is a server for the prediction of these short linear motifs within a set

of proteins. Given a set of sequences sharing a common functional feature (e.g.

interaction partner or localization) the method finds statistically

over-represented motifs likely to be responsible for it. The input sequences are

first passed through a set of filters to remove regions unlikely to contain

instances of linear motifs. Motifs are then found in the remaining sequence and

ranked according to a statistic that measure over-representation and conservation

across homologues in related species. The results are displayed via a visual

interface for easy perusal. The server is available at http://dilimot.embl.de.

DOI: 10.1093/nar/gkl159

PMCID: PMC1538856

PMID: 16845024 [Indexed for MEDLINE]

2927. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W325-9.

kinDOCK: a tool for comparative docking of protein kinase ligands.

Martin L(1), Catherinot V, Labesse G.

Author information:

(1)Atelier de Bio- et Chimie-Informatique Structurale, Centre de Biochimie

Structurale, CNRS UMR5048, Montpellier, France.

KinDOCK is a new web server for the analysis of ATP-binding sites of protein

kinases. This characterization is based on the docking of ligands already

co-crystallized with other protein kinases. A structural library of protein

kinase-ligand complexes has been extracted from the Protein Data Bank (PDB). This

library can provide both potential ligands and their putative binding orientation

for a given protein kinase. After protein-protein structural superposition, the

ligands are transferred from the template complexes to the target protein kinase.

The resulting complexes are evaluated using the program SCORE to compute a

theoretical affinity. They can be dynamically visualized to allow a rapid mapping

of important steric clashes and potential substitutions relevant for specificity

and affinity. These characteristics allow a quick characterization of protein

kinase active sites including conformation changes potentially required to

accommodate particular ligands. Additionally, promising pharmacophores can be

identified in the focussed library. These features will help to rationalize or

optimize virtual screening (VS) on larger chemical compound libraries. The server

and its documentation are freely available at http://abcis.cbs.cnrs.fr/kindock/.

DOI: 10.1093/nar/gkl211

PMCID: PMC1538843

PMID: 16845019 [Indexed for MEDLINE]

2928. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W32-7.

PROFEAT: a web server for computing structural and physicochemical features of

proteins and peptides from amino acid sequence.

Li ZR(1), Lin HH, Han LY, Jiang L, Chen X, Chen YZ.

Author information:

(1)Bioinformatics and Drug Design Group, Department of Computational Science,

National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore

117543.

Sequence-derived structural and physicochemical features have frequently been

used in the development of statistical learning models for predicting proteins

and peptides of different structural, functional and interaction profiles.

PROFEAT (Protein Features) is a web server for computing commonly-used structural

and physicochemical features of proteins and peptides from amino acid sequence.

It computes six feature groups composed of ten features that include 51

descriptors and 1447 descriptor values. The computed features include amino acid

composition, dipeptide composition, normalized Moreau-Broto autocorrelation,

Moran autocorrelation, Geary autocorrelation, sequence-order-coupling number,

quasi-sequence-order descriptors and the composition, transition and distribution

of various structural and physicochemical properties. In addition, it can also

compute previous autocorrelations descriptors based on user-defined properties.

Our computational algorithms were extensively tested and the computed protein

features have been used in a number of published works for predicting proteins of

functional classes, protein-protein interactions and MHC-binding peptides.

PROFEAT is accessible at http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi.

DOI: 10.1093/nar/gkl305

PMCID: PMC1538821

PMID: 16845018 [Indexed for MEDLINE]

2929. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W315-9.

TSEMA: interactive prediction of protein pairings between interacting families.

Izarzugaza JM(1), Juan D, Pons C, Ranea JA, Valencia A, Pazos F.

Author information:

(1)National Center for Biotechnology (CNB-CSIC) C/Darwin, 3 Cantoblanco, 28049

Madrid, Spain.

An entire family of methodologies for predicting protein interactions is based on

the observed fact that families of interacting proteins tend to have similar

phylogenetic trees due to co-evolution. One application of this concept is the

prediction of the mapping between the members of two interacting protein families

(which protein within one family interacts with which protein within the other).

The idea is that the real mapping would be the one maximizing the similarity

between the trees. Since the exhaustive exploration of all possible mappings is

not feasible for large families, current approaches use heuristic techniques

which do not ensure the best solution to be found. This is why it is important to

check the results proposed by heuristic techniques and to manually explore other

solutions. Here we present TSEMA, the server for efficient mapping assessment.

This system calculates an initial mapping between two families of proteins based

on a Monte Carlo approach and allows the user to interactively modify it based on

performance figures and/or specific biological knowledge. All the explored

mappings are graphically shown over a representation of the phylogenetic trees.

The system is freely available at http://pdg.cnb.uam.es/TSEMA. Standalone

versions of the software behind the interface are available upon request from the

authors.

DOI: 10.1093/nar/gkl112

PMCID: PMC1538787

PMID: 16845017 [Indexed for MEDLINE]

2930. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W310-4.

GRAMM-X public web server for protein-protein docking.

Tovchigrechko A(1), Vakser IA.

Author information:

(1)Center for Bioinformatics, The University of Kansas, 2030 Becker Drive,

Lawrence, KS 66047, USA. andrey@ku.edu

Protein docking software GRAMM-X and its web interface

(http://vakser.bioinformatics.ku.edu/resources/gramm/grammx) extend the original

GRAMM Fast Fourier Transformation methodology by employing smoothed potentials,

refinement stage, and knowledge-based scoring. The web server frees users from

complex installation of database-dependent parallel software and maintaining

large hardware resources needed for protein docking simulations. Docking problems

submitted to GRAMM-X server are processed by a 320 processor Linux cluster. The

server was extensively tested by benchmarking, several months of public use, and

participation in the CAPRI server track.

DOI: 10.1093/nar/gkl206

PMCID: PMC1538913

PMID: 16845016 [Indexed for MEDLINE]

2931. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W303-9.

Protemot: prediction of protein binding sites with automatically extracted

geometrical templates.

Chang DT(1), Weng YZ, Lin JH, Hwang MJ, Oyang YJ.

Author information:

(1)Department of Computer Science and Information Engineering, National Taiwan

University, Taipei, 106, Taiwan, Republic of China. darby@csie.ntu.edu.tw

Geometrical analysis of protein tertiary substructures has been an effective

approach employed to predict protein binding sites. This article presents the

Protemot web server that carries out prediction of protein binding sites based on

the structural templates automatically extracted from the crystal structures of

protein-ligand complexes in the PDB (Protein Data Bank). The automatic extraction

mechanism is essential for creating and maintaining a comprehensive template

library that timely accommodates to the new release of PDB as the number of

entries continues to grow rapidly. The design of Protemot is also distinctive by

the mechanism employed to expedite the analysis process that matches the tertiary

substructures on the contour of the query protein with the templates in the

library. This expediting mechanism is essential for providing reasonable response

time to the user as the number of entries in the template library continues to

grow rapidly due to rapid growth of the number of entries in PDB. This article

also reports the experiments conducted to evaluate the prediction power delivered

by the Protemot web server. Experimental results show that Protemot can deliver a

superior prediction power than a web server based on a manually curated template

library with insufficient quantity of entries.AVAILABILITY:

http://protemot.csie.ntu.edu.tw/step1.cgi

http://bioinfo.mc.ntu.edu.tw/protemot/step1.cgi.

DOI: 10.1093/nar/gkl344

PMCID: PMC1538868

PMID: 16845015 [Indexed for MEDLINE]

2932. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W3-5.

A compilation of molecular biology web servers: 2006 update on the Bioinformatics

Links Directory.

Fox JA(1), McMillan S, Ouellette BF.

Author information:

(1)UBC Bioinformatics Centre, University of British Columbia, Michael Smith

Laboratories, 2185 East Mall, Vancouver, BC, Canada V6T 1Z4.

The Bioinformatics Links Directory is a public online resource that lists the

servers published in this and all previously published Nucleic Acids Research Web

Server issues together with other useful tools, databases and resources for

bioinformatics and molecular biology research. This rich directory of tools and

websites can be browsed and searched with all listed links freely accessible to

the public. The 2006 update includes the 149 websites highlighted in the July

2006 issue of Nucleic Acids Research and brings the total number of servers

listed in the Bioinformatics Links Directory to over 1000 links. To aid

navigation through this growing resource, all link entries contain a brief

synopsis, a citation list and are classified by function in descriptive

biological categories. The most up-to-date version of this actively maintained

listing of bioinformatics resources is available at the Bioinformatics Links

Directory website, http://bioinformatics.ubc.ca/resources/links\_directory/. A

complete list of all links listed in this Nucleic Acids Research 2006 Web Server

issue can be accessed online at

http://bioinformatics.ubc.ca/resources/links\_directory/narweb2006/. The 2006

update of the Bioinformatics Links Directory, which includes the Web Server list

and summaries, is also available online at the Nucleic Acids Research website,

http://nar.oupjournals.org/.

DOI: 10.1093/nar/gkl379

PMCID: PMC1538876

PMID: 16845014 [Indexed for MEDLINE]

2933. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W273-9.

BAGEL: a web-based bacteriocin genome mining tool.

de Jong A(1), van Hijum SA, Bijlsma JJ, Kok J, Kuipers OP.

Author information:

(1)Department of Molecular Genetics, University of Groningen, Groningen

Biomolecular Sciences and Biotechnology Institute, PO Box 14, 9750 AA Haren, The

Netherlands.

A common problem in the annotation of open reading frames (ORFs) is the

identification of genes that are functionally similar but have limited or no

sequence homology. This is particularly the case for bacteriocins, a very diverse

group of antimicrobial peptides produced by bacteria and usually encoded by

small, poorly conserved ORFs. ORFs surrounding bacteriocin genes are often

biosynthetic genes. This information can be used to locate putative structural

bacteriocin genes. Here, we describe BAGEL, a web server that identifies putative

bacteriocin ORFs in a DNA sequence using novel, knowledge-based bacteriocin

databases and motif databases. Many bacteriocins are encoded by small genes that

are often omitted in the annotation process of bacterial genomes. Thus, we have

implemented ORF detection using a number of published ORF prediction tools. In

addition, BAGEL takes into account the genomic context, i.e. for each potential

bacteriocin-encoding ORF, the sequence of the surrounding region on the genome is

analyzed for genes that might encode proteins involved in biosynthesis,

transport, regulation and/or immunity. These innovations make BAGEL unique in its

ability to detect putative bacteriocin gene clusters in (new) bacterial genomes.

BAGEL is freely accessible at:

http://bioinformatics.biol.rug.nl/websoftware/bagel.

DOI: 10.1093/nar/gkl237

PMCID: PMC1538908

PMID: 16845009 [Indexed for MEDLINE]

2934. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W264-6.

DyNAVacS: an integrative tool for optimized DNA vaccine design.

Harish N(1), Gupta R, Agarwal P, Scaria V, Pillai B.

Author information:

(1)Indian Institute of Technology, Madras, Chennai.

DNA vaccines have slowly emerged as keystones in preventive immunology due to

their versatility in inducing both cell-mediated as well as humoral immune

responses. The design of an efficient DNA vaccine, involves choice of a suitable

expression vector, ensuring optimal expression by codon optimization, engineering

CpG motifs for enhancing immune responses and providing additional sequence

signals for efficient translation. DyNAVacS is a web-based tool created for rapid

and easy design of DNA vaccines. It follows a step-wise design flow, which guides

the user through the various sequential steps in the design of the vaccine.

Further, it allows restriction enzyme mapping, design of primers spanning user

specified sequences and provides information regarding the vectors currently used

for generation of DNA vaccines. The web version uses Apache HTTP server. The

interface was written in HTML and utilizes the Common Gateway Interface scripts

written in PERL for functionality. DyNAVacS is an integrated tool consisting of

user-friendly programs, which require minimal information from the user. The

software is available free of cost, as a web based application at URL:

http://miracle.igib.res.in/dynavac/.

DOI: 10.1093/nar/gkl242

PMCID: PMC1538838

PMID: 16845007 [Indexed for MEDLINE]

2935. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W258-63.

MODi: a powerful and convenient web server for identifying multiple

post-translational peptide modifications from tandem mass spectra.

Kim S(1), Na S, Sim JW, Park H, Jeong J, Kim H, Seo Y, Seo J, Lee KJ, Paek E.

Author information:

(1)Research Center for Computer Technology, Hanyang University, Seoul, 133-791,

Korea.

MOD(i) (http://modi.uos.ac.kr/modi/) is a powerful and convenient web service

that facilitates the interpretation of tandem mass spectra for identifying

post-translational modifications (PTMs) in a peptide. It is powerful in that it

can interpret a tandem mass spectrum even when hundreds of modification types are

considered and the number of potential PTMs in a peptide is large, in contrast to

most of the methods currently available for spectra interpretation that limit the

number of PTM sites and types being used for PTM analysis. For example, using

MOD(i), one can consider for analysis both the entire PTM list published on the

unimod webpage (http://www.unimod.org) and user-defined PTMs simultaneously, and

one can also identify multiple PTM sites in a spectrum. MOD(i) is convenient in

that it can take various input file formats such as .mzXML, .dta, .pkl and .mgf

files, and it is equipped with a graphical tool called MassPective developed to

display MOD(i)'s output in a user-friendly manner and helps users understand

MOD(i)'s output quickly. In addition, one can perform manual de novo sequencing

using MassPective.

DOI: 10.1093/nar/gkl245

PMCID: PMC1538808

PMID: 16845006 [Indexed for MEDLINE]

2936. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W254-7.

SUMOsp: a web server for sumoylation site prediction.

Xue Y(1), Zhou F, Fu C, Xu Y, Yao X.

Author information:

(1)Laboratory of Cellular Dynamics, Hefei National Laboratory for Physical

Sciences, and the University of Science and Technology of China, Hefei, China

230027.

Systematic dissection of the sumoylation proteome is emerging as an appealing but

challenging research topic because of the significant roles sumoylation plays in

cellular dynamics and plasticity. Although several proteome-scale analyzes have

been performed to delineate potential sumoylatable proteins, the bona fide

sumoylation sites still remain to be identified. Previously, we carried out a

genome-wide analysis of the SUMO substrates in human nucleus using the putative

motif psi-K-X-E and evolutionary conservation. However, a highly specific

predictor for in silico prediction of sumoylation sites in any individual

organism is still urgently needed to guide experimental design. In this work, we

present a computational system SUMOsp--SUMOylation Sites Prediction, based on a

manually curated dataset, integrating the results of two methods, GPS and MotifX,

which were originally designed for phosphorylation site prediction. SUMOsp offers

at least as good prediction performance as the only available method, SUMOplot,

on a very large test set. We expect that the prediction results of SUMOsp

combined with experimental verifications will propel our understanding of

sumoylation mechanisms to a new level. SUMOsp has been implemented on a freely

accessible web server at: http://bioinformatics.lcd-ustc.org/sumosp/.

DOI: 10.1093/nar/gkl207

PMCID: PMC1538802

PMID: 16845005 [Indexed for MEDLINE]

2937. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W249-53.

MeMo: a web tool for prediction of protein methylation modifications.

Chen H(1), Xue Y, Huang N, Yao X, Sun Z.

Author information:

(1)Institute of Bioinformatics and Systems Biology, MOE Key Laboratory of

Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology,

Department of Biological Sciences and Biotechnology, Tsinghua University,

Beijing, China 100084.

Protein methylation is an important and reversible post-translational

modification of proteins (PTMs), which governs cellular dynamics and plasticity.

Experimental identification of the methylation site is labor-intensive and often

limited by the availability of reagents, such as methyl-specific antibodies and

optimization of enzymatic reaction. Computational analysis may facilitate the

identification of potential methylation sites with ease and provide insight for

further experimentation. Here we present a novel protein methylation prediction

web server named MeMo, protein methylation modification prediction, implemented

in Support Vector Machines (SVMs). Our present analysis is primarily focused on

methylation on lysine and arginine, two major protein methylation sites. However,

our computational platform can be easily extended into the analyses of other

amino acids. The accuracies for prediction of protein methylation on lysine and

arginine have reached 67.1 and 86.7%, respectively. Thus, the MeMo system is a

novel tool for predicting protein methylation and may prove useful in the study

of protein methylation function and dynamics. The MeMo web server is available

at: http://www.bioinfo.tsinghua.edu.cn/~tigerchen/memo.html.

DOI: 10.1093/nar/gkl233

PMCID: PMC1538891

PMID: 16845004 [Indexed for MEDLINE]

2938. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W24-31.

oGNM: online computation of structural dynamics using the Gaussian Network Model.

Yang LW(1), Rader AJ, Liu X, Jursa CJ, Chen SC, Karimi HA, Bahar I.

Author information:

(1)Department of Computational Biology, School of Medicine, University of

Pittsburgh, Pittsburgh, PA 15213, USA.

An assessment of the equilibrium dynamics of biomolecular systems, and in

particular their most cooperative fluctuations accessible under native state

conditions, is a first step towards understanding molecular mechanisms relevant

to biological function. We present a web-based system, oGNM that enables users to

calculate online the shape and dispersion of normal modes of motion for proteins,

oligonucleotides and their complexes, or associated biological units, using the

Gaussian Network Model (GNM). Computations with the new engine are 5-6 orders of

magnitude faster than those using conventional normal mode analyses. Two cases

studies illustrate the utility of oGNM. The first shows that the thermal

fluctuations predicted for 1250 non-homologous proteins correlate well with X-ray

crystallographic data over a broad range [7.3-15 A] of inter-residue interaction

cutoff distances and the correlations improve with increasing observation

temperatures. The second study, focused on 64 oligonucleotides and

oligonucleotide-protein complexes, shows that good agreement with experiments is

achieved by representing each nucleotide by three GNM nodes (as opposed to

one-node-per-residue in proteins) along with uniform interaction ranges for all

components of the complexes. These results open the way to a rapid assessment of

the dynamics of DNA/RNA-containing complexes. The server can be accessed at

http://ignm.ccbb.pitt.edu/GNM\_Online\_Calculation.htm.

DOI: 10.1093/nar/gkl084

PMCID: PMC1538811

PMID: 16845002 [Indexed for MEDLINE]

2939. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W235-8.

RosettaDesign server for protein design.

Liu Y(1), Kuhlman B.

Author information:

(1)Department of Biochemistry and Biophysics, University of North Carolina,

Chapel Hill, NC 27599, USA.

The RosettaDesign server identifies low energy amino acid sequences for target

protein structures (http://rosettadesign.med.unc.edu). The client provides the

backbone coordinates of the target structure and specifies which residues to

design. The server returns to the client the sequences, coordinates and energies

of the designed proteins. The simulations are performed using the design module

of the Rosetta program (RosettaDesign). RosettaDesign uses Monte Carlo

optimization with simulated annealing to search for amino acids that pack well on

the target structure and satisfy hydrogen bonding potential. RosettaDesign has

been experimentally validated and has been used previously to stabilize naturally

occurring proteins and design a novel protein structure.

DOI: 10.1093/nar/gkl163

PMCID: PMC1538902

PMID: 16845000 [Indexed for MEDLINE]

2940. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W231-4.

HARMONY: a server for the assessment of protein structures.

Pugalenthi G(1), Shameer K, Srinivasan N, Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences (TIFR), GKVK Campus, Bellary Road,

Bangalore 560 065, India.

Protein structure validation is an important step in computational modeling and

structure determination. Stereochemical assessment of protein structures examine

internal parameters such as bond lengths and Ramachandran (varphi,psi) angles.

Gross structure prediction methods such as inverse folding procedure and

structure determination especially at low resolution can sometimes give rise to

models that are incorrect due to assignment of misfolds or mistracing of electron

density maps. Such errors are not reflected as strain in internal parameters.

HARMONY is a procedure that examines the compatibility between the sequence and

the structure of a protein by assigning scores to individual residues and their

amino acid exchange patterns after considering their local environments. Local

environments are described by the backbone conformation, solvent accessibility

and hydrogen bonding patterns. We are now providing HARMONY through a web server

such that users can submit their protein structure files and, if required, the

alignment of homologous sequences. Scores are mapped on the structure for

subsequent examination that is useful to also recognize regions of possible local

errors in protein structures. HARMONY server is located at

http://caps.ncbs.res.in/harmony/

DOI: 10.1093/nar/gkl314

PMCID: PMC1538917

PMID: 16844999 [Indexed for MEDLINE]

2941. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W219-24.

TarFisDock: a web server for identifying drug targets with docking approach.

Li H(1), Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J, Wang

X, Jiang H.

Author information:

(1)Drug Discovery and Design Center, State Key Laboratory of Drug Research,

Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai

201203, China.

TarFisDock is a web-based tool for automating the procedure of searching for

small molecule-protein interactions over a large repertoire of protein

structures. It offers PDTD (potential drug target database), a target database

containing 698 protein structures covering 15 therapeutic areas and a reverse

ligand-protein docking program. In contrast to conventional ligand-protein

docking, reverse ligand-protein docking aims to seek potential protein targets by

screening an appropriate protein database. The input file of this web server is

the small molecule to be tested, in standard mol2 format; TarFisDock then

searches for possible binding proteins for the given small molecule by use of a

docking approach. The ligand-protein interaction energy terms of the program DOCK

are adopted for ranking the proteins. To test the reliability of the TarFisDock

server, we searched the PDTD for putative binding proteins for vitamin E and

4H-tamoxifen. The top 2 and 10% candidates of vitamin E binding proteins

identified by TarFisDock respectively cover 30 and 50% of reported targets

verified or implicated by experiments; and 30 and 50% of experimentally confirmed

targets for 4H-tamoxifen appear amongst the top 2 and 5% of the TarFisDock

predicted candidates, respectively. Therefore, TarFisDock may be a useful tool

for target identification, mechanism study of old drugs and probes discovered

from natural products. TarFisDock and PDTD are available at

http://www.dddc.ac.cn/tarfisdock/.

DOI: 10.1093/nar/gkl114

PMCID: PMC1538869

PMID: 16844997 [Indexed for MEDLINE]

2942. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W210-3.

pTARGET: a web server for predicting protein subcellular localization.

Guda C(1).

Author information:

(1)Gen\*NY\*sis Center for Excellence in Cancer Genomics and Department of

Epidemiology and Biostatistics, University at Albany, State University of New

York, 1 Discovery drive, Rensselaer, NY 12144-3456, USA.

The pTARGET web server enables prediction of nine distinct protein subcellular

localizations in eukaryotic non-plant species. Predictions are made using a new

algorithm [C. Guda and S. Subramaniam (2005) pTARGET [corrected] a new method for

predicting protein subcellular localization in eukaryotes. Bioinformatics, 21,

3963-3969], which is primarily based on the occurrence patterns of

location-specific protein functional domains in different subcellular locations.

We have implemented a relational database, PreCalcDB, to store pre-computed

prediction results for all eukaryotic non-plant protein sequences in the public

domain that includes about 770,000 entries. Queries can be made by entering

protein sequences or by uploading a file containing up to 5000 protein sequences

in FASTA format. Prediction results for queries with matching entries in the

PreCalcDB will be retrieved instantly; while for the missing ones new predictions

will be computed and sent by email. Pre-computed predictions can also be

downloaded for complete proteomes of Saccharomyces cerevisiae, Caenorhabditis

elegans, Drosophila, Mus musculus and Homo sapiens. The server, its documentation

and the data are accessible from http://bioinformatics.albany.edu/~ptarget.

DOI: 10.1093/nar/gkl093

PMCID: PMC1538910

PMID: 16844995 [Indexed for MEDLINE]

2943. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W202-9.

AlgPred: prediction of allergenic proteins and mapping of IgE epitopes.

Saha S(1), Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector-39A,

Chandigarh, India.

In this study a systematic attempt has been made to integrate various approaches

in order to predict allergenic proteins with high accuracy. The dataset used for

testing and training consists of 578 allergens and 700 non-allergens obtained

from A. K. Bjorklund, D. Soeria-Atmadja, A. Zorzet, U. Hammerling and M. G.

Gustafsson (2005) Bioinformatics, 21, 39-50. First, we developed methods based on

support vector machine using amino acid and dipeptide composition and achieved an

accuracy of 85.02 and 84.00%, respectively. Second, a motif-based method has been

developed using MEME/MAST software that achieved sensitivity of 93.94 with 33.34%

specificity. Third, a database of known IgE epitopes was searched and this

predicted allergenic proteins with 17.47% sensitivity at specificity of 98.14%.

Fourth, we predicted allergenic proteins by performing BLAST search against

allergen representative peptides. Finally hybrid approaches have been developed,

which combine two or more than two approaches. The performance of all these

algorithms has been evaluated on an independent dataset of 323 allergens and on

101 725 non-allergens obtained from Swiss-Prot. A web server AlgPred has been

developed for the predicting allergenic proteins and for mapping IgE epitopes on

allergenic proteins (http://www.imtech.res.in/raghava/algpred/). AlgPred is

available at www.imtech.res.in/raghava/algpred/.

DOI: 10.1093/nar/gkl343

PMCID: PMC1538830

PMID: 16844994 [Indexed for MEDLINE]

2944. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W320-4.

PreBI: prediction of biological interfaces of proteins in crystals.

Tsuchiya Y(1), Kinoshita K, Ito N, Nakamura H.

Author information:

(1)Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka,

565-0871, Japan.

PreBI is a server that predicts biological interfaces in protein crystal

structures, according to the complementarity and the area of the interface. The

server accepts a coordinate file in the PDB format, and all of the possible

interfaces are generated automatically, according to the symmetry operations

given in the coordinate file. For all of the interfaces generated, the

complementarities of the electrostatic potential, hydrophobicity and shape of the

interfaces are analyzed, and the most probable biological interface is identified

according to the combination of the degree of complementarity derived from the

database analyses and the area of the interface. The results can be checked

through an interactive viewer, and the most probable complex can be downloaded as

atomic coordinates in the PDB format. PreBI is available at

http://pre-s.protein.osaka-u.ac.jp/~prebi/.

DOI: 10.1093/nar/gkl267

PMCID: PMC1538861

PMID: 16844993 [Indexed for MEDLINE]

2945. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W20-3.

PAST: fast structure-based searching in the PDB.

Täubig H(1), Buchner A, Griebsch J.

Author information:

(1)Efficient Algorithms Group, Department of Computer Science, Technische

Universität München, Boltzmannstrasse. 3, 85748 Garching, Germany.

PAST is a new web service providing fast structural queries of the Protein Data

Bank. The search engine is based on an adaptation of the generalized suffix tree

and relies on a translation- and rotation-invariant representation of the protein

backbone. The search procedure is completely independent of the amino acid

sequence of the polypeptide chains. The web service works best with, but is not

necessarily limited to, shorter fragments such as functional motifs-a task that

most other tools do not perform well. Usual query times are in the order of

seconds, allowing a truly interactive use. Unlike most established tools, PAST

does not prefilter the dataset or exclude parts of the search space based on

statistical reasoning. The server is freely available at http://past.in.tum.de/.

DOI: 10.1093/nar/gkl273

PMCID: PMC1538836

PMID: 16844992 [Indexed for MEDLINE]

2946. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W198-201.

REMUS: a tool for identification of unique peptide segments as epitopes.

Pai TW(1), Chang MD, Tzou WS, Su BH, Wu PC, Chang HT, Chou WI.

Author information:

(1)Department of Computer Science and Engineering, National Taiwan Ocean

University, Keelung, Taiwan, Republic of China. twp@mail.ntou.edu.tw

We provide a 'R(E)MUS' (reinforced merging techniques for unique peptide

segments) web server for identification of the locations and compositions of

unique peptide segments from a set of protein family sequences. Different levels

of uniqueness are determined according to substitutional relationship in the

amino acids, frequency of appearance and biological properties such as priority

for serving as candidates for epitopes where antibodies recognize. R(E)MUS also

provides interactive visualization of 3D structures for allocation and comparison

of the identified unique peptide segments. Accuracy of the algorithm was found to

be 70% in terms of mapping a unique peptide segment as an epitope. The R(E)MUS

web server is available at http://biotools.cs.ntou.edu.tw/REMUS and the PC

version software can be freely downloaded either at

http://bioinfo.life.nthu.edu.tw/REMUS or

http://spider.cs.ntou.edu.tw/BioTools/REMUS. User guide and working examples for

PC version are available at

http://spider.cs.ntou.edu.tw/BioTools/REMUS-DOCS.html, and details of the

proposed algorithm can be referred to the documents as described previously [H.

T. Chang, T. W. Pai, T. C. Fan, B. H. Su, P. C. Wu, C. Y. Tang, C. T. Chang, S.

H. Liu and M. D. T. Chang (2006) BMC Bioinformatics, 7, 38 and T. W. Pai, B. H.

Su, P. C. Wu, M. D. T. Chang, H. T. Chang, T. C. Fan and S. H. Liu (2006) J.

Bioinform. Comput. Biol., 4, 75-92].

DOI: 10.1093/nar/gkl188

PMCID: PMC1538771

PMID: 16844991 [Indexed for MEDLINE]

2947. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W194-7.

SVMHC: a server for prediction of MHC-binding peptides.

Dönnes P(1), Kohlbacher O.

Author information:

(1)Division for Simulation of Biological Systems, WSI/ZBIT, Eberhard Karls

University, Tübingen Sand 14, D-72076 Tübingen, Germany.

doennes@informatik.uni-tuebingen.de

Identification of MHC-binding peptides is a prerequisite in rational design of

T-cell based peptide vaccines. During the past decade a number of computational

approaches have been introduced for the prediction of MHC-binding peptides,

efficiently reducing the number of candidate binders that need to be

experimentally verified. Here the SVMHC server for prediction of both MHC class I

and class II binding peptides is presented. SVMHC offers fast analysis of a wide

range of alleles and prediction results are given in several comprehensive

formats. The server can be used to find the most likely binders in a protein

sequence and to investigate the effects of single nucleotide polymorphisms in

terms of MHC-peptide binding. The SVMHC server is accessible at

http://www-bs.informatik.uni-tuebingen.de/SVMHC/.

DOI: 10.1093/nar/gkl284

PMCID: PMC1538857

PMID: 16844990 [Indexed for MEDLINE]

2948. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W189-93.

transFold: a web server for predicting the structure and residue contacts of

transmembrane beta-barrels.

Waldispühl J(1), Berger B, Clote P, Steyaert JM.

Author information:

(1)Department of Biology, Boston College, 140 Commonwealth Avenue, Chestnut Hill,

MA 02467, USA.

Transmembrane beta-barrel (TMB) proteins are embedded in the outer membrane of

Gram-negative bacteria, mitochondria and chloroplasts. The cellular location and

functional diversity of beta-barrel outer membrane proteins makes them an

important protein class. At the present time, very few non-homologous TMB

structures have been determined by X-ray diffraction because of the experimental

difficulty encountered in crystallizing transmembrane (TM) proteins. The

transFold web server uses pairwise inter-strand residue statistical potentials

derived from globular (non-outer-membrane) proteins to predict the supersecondary

structure of TMB. Unlike all previous approaches, transFold does not use machine

learning methods such as hidden Markov models or neural networks; instead,

transFold employs multi-tape S-attribute grammars to describe all potential

conformations, and then applies dynamic programming to determine the global

minimum energy supersecondary structure. The transFold web server not only

predicts secondary structure and TMB topology, but is the only method which

additionally predicts the side-chain orientation of transmembrane beta-strand

residues, inter-strand residue contacts and TM beta-strand inclination with

respect to the membrane. The program transFold currently outperforms all other

methods for accuracy of beta-barrel structure prediction. Available at

http://bioinformatics.bc.edu/clotelab/transFold.

DOI: 10.1093/nar/gkl205

PMCID: PMC1538872

PMID: 16844989 [Indexed for MEDLINE]

2949. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W186-8.

PROFtmb: a web server for predicting bacterial transmembrane beta barrel

proteins.

Bigelow H(1), Rost B.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 630 West 168th Street, New York, NY 10032, USA. hrbigelow@gmail.com

PROFtmb predicts transmembrane beta-barrel (TMB) proteins in Gram-negative

bacteria. For each query protein, PROFtmb provides both a Z-value indicating that

the protein actually contains a membrane barrel, and a four-state per-residue

labeling of upward- and downward-facing strands, periplasmic hairpins and

extracellular loops. While most users submit individual proteins known to contain

TMBs, some groups submit entire proteomes to screen for potential TMBs. Response

time is about 4 min for a 500-residue protein. PROFtmb is a profile-based Hidden

Markov Model (HMM) with an architecture mirroring the structure of TMBs. The

per-residue accuracy on the 8-fold cross-validated testing set is 86% while

whole-protein discrimination accuracy was 70 at 60% coverage. The PROFtmb web

server includes all source code, training data and whole-proteome predictions

from 78 Gram-negative bacterial genomes and is available freely and without

registration at http://rostlab.org/services/proftmb.

DOI: 10.1093/nar/gkl262

PMCID: PMC1538807

PMID: 16844988 [Indexed for MEDLINE]

2950. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W182-5.

DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine

classification.

Ferrè F(1), Clote P.

Author information:

(1)Department of Biology, Boston College, Chestnut Hill, MA, USA.

DiANNA is a recent state-of-the-art artificial neural network and web server,

which determines the cysteine oxidation state and disulfide connectivity of a

protein, given only its amino acid sequence. Version 1.0 of DiANNA uses a

feed-forward neural network to determine which cysteines are involved in a

disulfide bond, and employs a novel architecture neural network to predict which

half-cystines are covalently bound to which other half-cystines. In version 1.1

of DiANNA, described here, we extend functionality by applying a support vector

machine with spectrum kernel for the cysteine classification problem-to determine

whether a cysteine is reduced (free in sulfhydryl state), half-cystine (involved

in a disulfide bond) or bound to a metallic ligand. In the latter case, DiANNA

predicts the ligand among iron, zinc, cadmium and carbon. Available at:

http://bioinformatics.bc.edu/clotelab/DiANNA/.

DOI: 10.1093/nar/gkl189

PMCID: PMC1538812

PMID: 16844987 [Indexed for MEDLINE]

2951. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W177-81.

DISULFIND: a disulfide bonding state and cysteine connectivity prediction server.

Ceroni A(1), Passerini A, Vullo A, Frasconi P.

Author information:

(1)Machine Learning and Neural Networks Group, Università degli Studi di Firenze,

Dipartimento di Sistemi e Informatica, Via di Santa Marta 3, 50139 Firenze,

Italy.

DISULFIND is a server for predicting the disulfide bonding state of cysteines and

their disulfide connectivity starting from sequence alone. Optionally, disulfide

connectivity can be predicted from sequence and a bonding state assignment given

as input. The output is a simple visualization of the assigned bonding state

(with confidence degrees) and the most likely connectivity patterns. The server

is available at http://disulfind.dsi.unifi.it/.

DOI: 10.1093/nar/gkl266

PMCID: PMC1538823

PMID: 16844986 [Indexed for MEDLINE]

2952. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W173-6.

ArchPRED: a template based loop structure prediction server.

Fernandez-Fuentes N(1), Zhai J, Fiser A.

Author information:

(1)Department of Biochemistry and Seaver Foundation Center for Bioinformatics,

Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461,

USA.

ArchPRED server (http://www.fiserlab.org/servers/archpred) implements a novel

fragment-search based method for predicting loop conformations. The inputs to the

server are the atomic coordinates of the query protein and the position of the

loop. The algorithm selects candidate loop fragments from a regularly updated

loop library (Search Space) by matching the length, the types of bracing

secondary structures of the query and by satisfying the geometrical restraints

imposed by the stem residues. Subsequently, candidate loops are inserted in the

query protein framework where their side chains are rebuilt and their fit is

assessed by the root mean square deviation (r.m.s.d.) of stem regions and by the

number of rigid body clashes with the environment. In the final step remaining

candidate loops are ranked by a Z-score that combines information on sequence

similarity and fit of predicted and observed [/psi] main chain dihedral angle

propensities. The final loop conformation is built in the protein structure and

annealed in the environment using conjugate gradient minimization. The prediction

method was benchmarked on artificially prepared search datasets where all trivial

sequence similarities on the SCOP superfamily level were removed. Under these

conditions it was possible to predict loops of length 4, 8 and 12 with coverage

of 98, 78 and 28% with at least of 0.22, 1.38 and 2.47 A of r.m.s.d. accuracy,

respectively. In a head to head comparison on loops extracted from freshly

deposited new protein folds the current method outperformed in a approximately

5:1 ratio an earlier developed database search method.

DOI: 10.1093/nar/gkl113

PMCID: PMC1538831

PMID: 16844985 [Indexed for MEDLINE]

2953. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W169-72.

PONGO: a web server for multiple predictions of all-alpha transmembrane proteins.

Amico M(1), Finelli M, Rossi I, Zauli A, Elofsson A, Viklund H, von Heijne G,

Jones D, Krogh A, Fariselli P, Luigi Martelli P, Casadio R.

Author information:

(1)BioDec Srl, via Calzavecchio, 20/2, I-40033 Casalecchio di Reno, (BO), Italy.

The annotation efforts of the BIOSAPIENS European Network of Excellence have

generated several distributed annotation systems (DAS) with the aim of

integrating Bioinformatics resources and annotating metazoan genomes

(http://www.biosapiens.info). In this context, the PONGO DAS server

(http://pongo.biocomp.unibo.it) provides the annotation on predictive basis for

the all-alpha membrane proteins in the human genome, not only through DAS

queries, but also directly using a simple web interface. In order to produce a

more comprehensive analysis of the sequence at hand, this annotation is carried

out with four selected and high scoring predictors: TMHMM2.0, MEMSAT, PRODIV and

ENSEMBLE1.0. The stored and pre-computed predictions for the human proteins can

be searched and displayed in a graphical view. However the web service allows the

prediction of the topology of any kind of putative membrane proteins, regardless

of the organism and more importantly with the same sequence profile for a given

sequence when required. Here we present a new web server that incorporates the

state-of-the-art topology predictors in a single framework, so that putative

users can interactively compare and evaluate four predictions simultaneously for

a given sequence. Together with the predicted topology, the server also displays

a signal peptide prediction determined with SPEP. The PONGO web server is

available at http://pongo.biocomp.unibo.it/pongo.

DOI: 10.1093/nar/gkl208

PMCID: PMC1538841

PMID: 16844984 [Indexed for MEDLINE]

2954. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W164-8.

Spritz: a server for the prediction of intrinsically disordered regions in

protein sequences using kernel machines.

Vullo A(1), Bortolami O, Pollastri G, Tosatto SC.

Author information:

(1)School of Computer Science and Informatics, University College Dublin,

Ireland.

Intrinsically disordered proteins have long stretches of their polypeptide chain,

which do not adopt a single native structure composed of stable secondary and

tertiary structure in the absence of binding partners. The prediction of

intrinsically disordered regions in proteins from sequence is increasingly

becoming of interest, as the presence of many such regions in the complete genome

sequences are discovered and important functional roles are associated with them.

We have developed a machine learning approach based on two support vector

machines (SVM) to discriminate disordered regions from sequence. The SVM are

trained and benchmarked on two sets, representing long and short disordered

regions. A preliminary version of Spritz was shown to perform consistently well

at the recent biannual CASP-6 experiment [Critical Assessment of Techniques for

Protein Structure Prediction (CASP), 2004]. The fully developed Spritz method is

freely available as a web server at http://distill.ucd.ie/spritz/ and

http://protein.cribi.unipd.it/spritz/.

DOI: 10.1093/nar/gkl166

PMCID: PMC1538873

PMID: 16844983 [Indexed for MEDLINE]

2955. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W158-63.

KemaDom: a web server for domain prediction using kernel machine with local

context.

Chen L(1), Wang W, Ling S, Jia C, Wang F.

Author information:

(1)Shanghai Key Laboratory of Intelligent Information Processing, Fudan

University, Shanghai, PR China.

Predicting domains of proteins is an important and challenging problem in

computational biology because of its significant role in understanding the

complexity of proteomes. Although many template-based prediction servers have

been developed, ab initio methods should be designed and further improved to be

the complementarity of the template-based methods. In this paper, we present a

novel domain prediction system KemaDom by ensembling three kernel machines with

the local context information among neighboring amino acids. KemaDom, an

alternative ab initio predictor, can achieve high performance in predicting the

number of domains in proteins. It is freely accessible at

http://www.iipl.fudan.edu.cn/lschen/kemadom.htm and

http://www.iipl.fudan.edu.cn/~lschen/kemadom.htm.

DOI: 10.1093/nar/gkl331

PMCID: PMC1538912

PMID: 16844982 [Indexed for MEDLINE]

2956. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W152-7.

(PS)2: protein structure prediction server.

Chen CC(1), Hwang JK, Yang JM.

Author information:

(1)Institute of Bioinformatics, National Chiao Tung University, Hsinchu, 30050,

Taiwan.

Protein structure prediction provides valuable insights into function, and

comparative modeling is one of the most reliable methods to predict 3D structures

directly from amino acid sequences. However, critical problems arise during the

selection of the correct templates and the alignment of query sequences

therewith. We have developed an automatic protein structure prediction server,

(PS)2, which uses an effective consensus strategy both in template selection,

which combines PSI-BLAST and IMPALA, and target-template alignment integrating

PSI-BLAST, IMPALA and T-Coffee. (PS)2 was evaluated for 47 comparative modeling

targets in CASP6 (Critical Assessment of Techniques for Protein Structure

Prediction). For the benchmark dataset, the predictive performance of (PS)2,

based on the mean GTD\_TS score, was superior to 10 other automatic servers. Our

method is based solely on the consensus sequence and thus is considerably faster

than other methods that rely on the additional structural consensus of templates.

Our results show that (PS)2, coupled with suitable consensus strategies and a new

similarity score, can significantly improve structure prediction. Our approach

should be useful in structure prediction and modeling. The (PS)2 is available

through the website at http://ps2.life.nctu.edu.tw/.

DOI: 10.1093/nar/gkl187

PMCID: PMC1538880

PMID: 16844981 [Indexed for MEDLINE]

2957. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W143-6.

Cascade PSI-BLAST web server: a remote homology search tool for relating protein

domains.

Bhadra R(1), Sandhya S, Abhinandan KR, Chakrabarti S, Sowdhamini R, Srinivasan N.

Author information:

(1)Molecular Biophysics Unit, Indian Institute of Science, 560 012, Bangalore,

India.

Owing to high evolutionary divergence, it is not always possible to identify

distantly related protein domains by sequence search techniques. Intermediate

sequences possess sequence features of more than one protein and facilitate

detection of remotely related proteins. We have demonstrated recently the

employment of Cascade PSI-BLAST where we perform PSI-BLAST for many

'generations', initiating searches from new homologues as well. Such a rigorous

propagation through generations of PSI-BLAST employs effectively the role of

intermediates in detecting distant similarities between proteins. This approach

has been tested on a large number of folds and its performance in detecting

superfamily level relationships is approximately 35% better than simple PSI-BLAST

searches. We present a web server for this search method that permits users to

perform Cascade PSI-BLAST searches against the Pfam, SCOP and SwissProt

databases. The URL for this server is

http://crick.mbu.iisc.ernet.in/~CASCADE/CascadeBlast.html.

DOI: 10.1093/nar/gkl157

PMCID: PMC1538780

PMID: 16844978 [Indexed for MEDLINE]

2958. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W137-42.

HHrep: de novo protein repeat detection and the origin of TIM barrels.

Söding J(1), Remmert M, Biegert A.

Author information:

(1)Department of Protein Evolution, Max-Planck-Institute for Developmental

Biology, Spemannstrasse 35, 72076 Tübingen, Germany.

johannes.soeding@tuebingen.mpg.de

HHrep is a web server for the de novo identification of repeats in protein

sequences, which is based on the pairwise comparison of profile hidden Markov

models (HMMs). Its main strength is its sensitivity, allowing it to detect highly

divergent repeat units in protein sequences whose repeats could as yet only be

detected from their structures. Examples include sequences with beta-propellor

fold, ferredoxin-like fold, double psi barrels or (betaalpha)8 (TIM) barrels. We

illustrate this with proteins from four superfamilies of TIM barrels by revealing

a clear 4- and 8-fold symmetry, which we detect solely from their sequences. This

symmetry might be the trace of an ancient origin through duplication of a

betaalphabetaalpha or betaalpha unit. HHrep can be accessed at

http://hhrep.tuebingen.mpg.de.

DOI: 10.1093/nar/gkl130

PMCID: PMC1538828

PMID: 16844977 [Indexed for MEDLINE]

2959. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W128-32.

3dSS: 3D structural superposition.

Sumathi K(1), Ananthalakshmi P, Roshan MN, Sekar K.

Author information:

(1)Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India.

3dSS is a web-based interactive computing server, primarily designed to aid

researchers, to superpose two or several 3D protein structures. In addition, the

server can be effectively used to find the invariant and common water molecules

present in the superposed homologous protein structures. The molecular

visualization tool RASMOL is interfaced with the server to visualize the

superposed 3D structures with the water molecules (invariant or common) in the

client machine. Furthermore, an option is provided to save the superposed 3D

atomic coordinates in the client machine. To perform the above, users need to

enter Protein Data Bank (PDB)-id(s) or upload the atomic coordinates in PDB

format. This server uses a locally maintained PDB anonymous FTP server that is

being updated weekly. This program can be accessed through our Bioinformatics web

server at the URL http://cluster.physics.iisc.ernet.in/3dss/ or

http://10.188.1.15/3dss/.

DOI: 10.1093/nar/gkl036

PMCID: PMC1538824

PMID: 16844975 [Indexed for MEDLINE]

2960. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W124-7.

ReadOut: structure-based calculation of direct and indirect readout energies and

specificities for protein-DNA recognition.

Ahmad S(1), Kono H, Araúzo-Bravo MJ, Sarai A.

Author information:

(1)Department of Bioscience and Bioinformatics, Kyushu Institute of Technology,

Iizuka 820 8502, Fukuoka, Japan.

Protein-DNA interactions play a central role in regulatory processes at the

genetic level. DNA-binding proteins recognize their targets by direct base-amino

acid interactions and indirect conformational energy contribution from DNA

deformations and elasticity. Knowledge-based approach based on the statistical

analysis of protein-DNA complex structures has been successfully used to

calculate interaction energies and specificities of direct and indirect readouts

in protein-DNA recognition. Here, we have implemented the method as a webserver,

which calculates direct and indirect readout energies and Z-scores, as a measure

of specificity, using atomic coordinates of protein-DNA complexes. This server is

freely available at http://gibk26.bse.kyutech.ac.jp/jouhou/readout/. The only

input to this webserver is the Protein Data Bank (PDB) style coordinate data of

atoms or the PDB code itself. The server returns total energy Z-scores, which

estimate the degree of sequence specificity of the protein-DNA complex. This

webserver is expected to be useful for estimating interaction energy and DNA

conformation energy, and relative contributions to the specificity from direct

and indirect readout. It may also be useful for checking the quality of

protein-DNA complex structures, and for engineering proteins and target DNAs.

DOI: 10.1093/nar/gkl104

PMCID: PMC1538882

PMID: 16844974 [Indexed for MEDLINE]

2961. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W119-23.

Protein Block Expert (PBE): a web-based protein structure analysis server using a

structural alphabet.

Tyagi M(1), Sharma P, Swamy CS, Cadet F, Srinivasan N, de Brevern AG, Offmann B.

Author information:

(1)Laboratoire de Biochimie et Génétique Moléculaire, Bioinformatics Team,

Université de La Réunion, BP 7151, 15 avenue René Cassin, 97715 Saint Denis

Messag Cedex 09, La Réunion, France.

Encoding protein 3D structures into 1D string using short structural prototypes

or structural alphabets opens a new front for structure comparison and analysis.

Using the well-documented 16 motifs of Protein Blocks (PBs) as structural

alphabet, we have developed a methodology to compare protein structures that are

encoded as sequences of PBs by aligning them using dynamic programming which uses

a substitution matrix for PBs. This methodology is implemented in the

applications available in Protein Block Expert (PBE) server. PBE addresses common

issues in the field of protein structure analysis such as comparison of proteins

structures and identification of protein structures in structural databanks that

resemble a given structure. PBE-T provides facility to transform any PDB file

into sequences of PBs. PBE-ALIGNc performs comparison of two protein structures

based on the alignment of their corresponding PB sequences. PBE-ALIGNm is a

facility for mining SCOP database for similar structures based on the alignment

of PBs. Besides, PBE provides an interface to a database (PBE-SAdb) of

preprocessed PB sequences from SCOP culled at 95% and of all-against-all pairwise

PB alignments at family and superfamily levels. PBE server is freely available at

http://bioinformatics.univ-reunion.fr/PBE/.

DOI: 10.1093/nar/gkl199

PMCID: PMC1538797

PMID: 16844973 [Indexed for MEDLINE]

2962. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W116-8.

CASTp: computed atlas of surface topography of proteins with structural and

topographical mapping of functionally annotated residues.

Dundas J(1), Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J.

Author information:

(1)Program in Bioinformatics, Department of Bioengineering, University of

Illinois at Chicago, Chicago, IL 60612, USA.

Cavities on a proteins surface as well as specific amino acid positioning within

it create the physicochemical properties needed for a protein to perform its

function. CASTp (http://cast.engr.uic.edu) is an online tool that locates and

measures pockets and voids on 3D protein structures. This new version of CASTp

includes annotated functional information of specific residues on the protein

structure. The annotations are derived from the Protein Data Bank (PDB),

Swiss-Prot, as well as Online Mendelian Inheritance in Man (OMIM), the latter

contains information on the variant single nucleotide polymorphisms (SNPs) that

are known to cause disease. These annotated residues are mapped to surface

pockets, interior voids or other regions of the PDB structures. We use a

semi-global pair-wise sequence alignment method to obtain sequence mapping

between entries in Swiss-Prot, OMIM and entries in PDB. The updated CASTp web

server can be used to study surface features, functional regions and specific

roles of key residues of proteins.

DOI: 10.1093/nar/gkl282

PMCID: PMC1538779

PMID: 16844972 [Indexed for MEDLINE]

2963. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W110-5.

TreeDet: a web server to explore sequence space.

Carro A(1), Tress M, de Juan D, Pazos F, Lopez-Romero P, del Sol A, Valencia A,

Rojas AM.

Author information:

(1)Protein Design Group, CNB-CSIC C/Darwin n.3, 28049, Madrid, Spain.

The TreeDet (Tree Determinant) Server is the first release of a system designed

to integrate results from methods that predict functional sites in protein

families. These methods take into account the relation between sequence

conservation and evolutionary importance. TreeDet fully analyses the space of

protein sequences in either user-uploaded or automatically generated multiple

sequence alignments. The methods implemented in the server represent three main

classes of methods for the detection of family-dependent conserved positions, a

tree-based method, a correlation based method and a method that employs a

principal component analyses coupled to a cluster algorithm. An additional method

is provided to highlight the reliability of the position in the alignments. The

server is available at http://www.pdg.cnb.uam.es/servers/treedet.

DOI: 10.1093/nar/gkl203

PMCID: PMC1538789

PMID: 16844971 [Indexed for MEDLINE]

2964. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W10-4.

FISH--family identification of sequence homologues using structure anchored

hidden Markov models.

Tångrot J(1), Wang L, Kågström B, Sauer UH.

Author information:

(1)Umeå Center for Molecular Pathogenesis UCMP, Umeå University Umeå, Sweden.

The FISH server is highly accurate in identifying the family membership of

domains in a query protein sequence, even in the case of very low sequence

identities to known homologues. A performance test using SCOP sequences and an

E-value cut-off of 0.1 showed that 99.3% of the top hits are to the correct

family saHMM. Matches to a query sequence provide the user not only with an

annotation of the identified domains and hence a hint to their function, but also

with probable 2D and 3D structures, as well as with pairwise and multiple

sequence alignments to homologues with low sequence identity. In addition, the

FISH server allows users to upload and search their own protein sequence

collection or to quarry public protein sequence data bases with individual

saHMMs. The FISH server can be accessed at http://babel.ucmp.umu.se/fish/.

DOI: 10.1093/nar/gkl330

PMCID: PMC1538871

PMID: 16844969 [Indexed for MEDLINE]

2965. Proteins. 2006 Jul 1;64(1):19-27.

Using evolutionary and structural information to predict DNA-binding sites on

DNA-binding proteins.

Kuznetsov IB(1), Gou Z, Li R, Hwang S.

Author information:

(1)Gen\*NY\*sis Center for Excellence in Cancer Genomics, Department of

Epidemiology and Biostatistics, University at Albany, Rensselaer, NewYork 12144,

USA. Ikuznetsov@albany.edu

Proteins that interact with DNA are involved in a number of fundamental

biological activities such as DNA replication, transcription, and repair. A

reliable identification of DNA-binding sites in DNA-binding proteins is important

for functional annotation, site-directed mutagenesis, and modeling protein-DNA

interactions. We apply Support Vector Machine (SVM), a supervised pattern

recognition method, to predict DNA-binding sites in DNA-binding proteins using

the following features: amino acid sequence, profile of evolutionary conservation

of sequence positions, and low-resolution structural information. We use a

rigorous statistical approach to study the performance of predictors that utilize

different combinations of features and how this performance is affected by

structural and sequence properties of proteins. Our results indicate that an SVM

predictor based on a properly scaled profile of evolutionary conservation in the

form of a position specific scoring matrix (PSSM) significantly outperforms a

PSSM-based neural network predictor. The highest accuracy is achieved by SVM

predictor that combines the profile of evolutionary conservation with

low-resolution structural information. Our results also show that knowledge-based

predictors of DNA-binding sites perform significantly better on proteins from

mainly-alpha structural class and that the performance of these predictors is

significantly correlated with certain structural and sequence properties of

proteins. These observations suggest that it may be possible to assign a

reliability index to the overall accuracy of the prediction of DNA-binding sites

in any given protein using its sequence and structural properties. A web-server

implementation of the predictors is freely available online at

http://lcg.rit.albany.edu/dp-bind/.

(c) 2006 Wiley-Liss, Inc.

DOI: 10.1002/prot.20977

PMID: 16568445 [Indexed for MEDLINE]

2966. BMC Bioinformatics. 2006 Jun 23;7:318.

MACSIMS: multiple alignment of complete sequences information management system.

Thompson JD(1), Muller A, Waterhouse A, Procter J, Barton GJ, Plewniak F, Poch O.

Author information:

(1)Laboratoire de Biologie et Genomique Structurales, Institut de Génétique et de

Biologie Moléculaire et Cellulaire, Illkirch, France. julie@igbmc.u-strasbg.fr

BACKGROUND: In the post-genomic era, systems-level studies are being performed

that seek to explain complex biological systems by integrating diverse resources

from fields such as genomics, proteomics or transcriptomics. New information

management systems are now needed for the collection, validation and analysis of

the vast amount of heterogeneous data available. Multiple alignments of complete

sequences provide an ideal environment for the integration of this information in

the context of the protein family.

RESULTS: MACSIMS is a multiple alignment-based information management program

that combines the advantages of both knowledge-based and ab initio sequence

analysis methods. Structural and functional information is retrieved

automatically from the public databases. In the multiple alignment, homologous

regions are identified and the retrieved data is evaluated and propagated from

known to unknown sequences with these reliable regions. In a large-scale

evaluation, the specificity of the propagated sequence features is estimated to

be >99%, i.e. very few false positive predictions are made. MACSIMS is then used

to characterise mutations in a test set of 100 proteins that are known to be

involved in human genetic diseases. The number of sequence features associated

with these proteins was increased by 60%, compared to the features available in

the public databases. An XML format output file allows automatic parsing of the

MACSIM results, while a graphical display using the JalView program allows manual

analysis.

CONCLUSION: MACSIMS is a new information management system that incorporates

detailed analyses of protein families at the structural, functional and

evolutionary levels. MACSIMS thus provides a unique environment that facilitates

knowledge extraction and the presentation of the most pertinent information to

the biologist. A web server and the source code are available at

http://bips.u-strasbg.fr/MACSIMS/.

DOI: 10.1186/1471-2105-7-318

PMCID: PMC1539025

PMID: 16792820 [Indexed for MEDLINE]

2967. BMC Bioinformatics. 2006 Jun 22;7:315.

PHY.FI: fast and easy online creation and manipulation of phylogeny color

figures.

Fredslund J(1).

Author information:

(1)Bioinformatics Research Center, Aarhus University, Høegh-Guldbergs Gade 10,

Bldg, 1090, DK-8000 Arhus C, Denmark. jakobf@birc.au.dk

BACKGROUND: The need to depict a phylogeny, or some other kind of abstract tree,

is very frequently experienced by researchers from a broad range of biological

and computational disciplines. Thousands of papers and talks include phylogeny

figures, and often during everyday work, one would like to quickly get a

graphical display of, e.g., the phylogenetic relationship between a set of

sequences as calculated by an alignment program such as ClustalW or the

phylogenetic package Phylip. A wealth of software tools capable of tree drawing

exists; most are comprehensive packages that also perform various types of

analysis, and hence they are available only for download and installing. Some

online tools exist, too.

RESULTS: This paper presents an online tool, PHY.FI, which encompasses all the

qualities of existing online programs and adds functionality to hopefully

eliminate the need for post-processing the phylogeny figure in some other

general-purpose graphics program. PHY.FI is versatile, easy-to-use and fast, and

supports comprehensive graphical control, several download image formats, and the

possibility of dynamically collapsing groups of nodes into named subtrees (e.g.

"Primates"). The user can create a color figure from any phylogeny, or other kind

of tree, represented in the widely used parenthesized Newick format.

CONCLUSION: PHY.FI is fast and easy to use, yet still offers full color control,

tree manipulation, and several image formats. It does not require any downloading

and installing, and thus any internet user regardless of computer skills, and

computer platform, can benefit from it. PHY.FI is free for all and is available

from this web address: http://cgi-www.daimi.au.dk/cgi-chili/phyfi/go.

DOI: 10.1186/1471-2105-7-315

PMCID: PMC1513607

PMID: 16792795 [Indexed for MEDLINE]

2968. BMC Struct Biol. 2006 Jun 22;6:13.

Prediction of transmembrane helix orientation in polytopic membrane proteins.

Adamian L(1), Liang J.

Author information:

(1)Department of Bioengineering, University of Illinois at Chicago, M/C 563, 835

S, Wolcott St, Chicago, IL 60612-7340, USA. larisa@uic.edu

BACKGROUND: Membrane proteins compose up to 30% of coding sequences within

genomes. However, their structure determination is lagging behind compared with

soluble proteins due to the experimental difficulties. Therefore, it is important

to develop reliable computational methods to predict structures of membrane

proteins.

RESULTS: We present a method for prediction of the TM helix orientation, which is

an essential step in ab initio modeling of membrane proteins. Our method is based

on a canonical model of the heptad repeat originally developed for coiled coils.

We identify the helical surface patches that interface with lipid molecules at an

accuracy of about 88% from the sequence information alone, using an empirical

scoring function LIPS (LIPid-facing Surface), which combines lipophilicity and

conservation of residues in the helix. We test and discuss results of prediction

of helix-lipid interfaces on 162 transmembrane helices from 18 polytopic membrane

proteins and present predicted orientations of TM helices in TRPV1 channel. We

also apply our method to two structures of homologous cytochrome b6f complexes

and find discrepancy in the assignment of TM helices from subunits PetG, PetN and

PetL. The results of LIPS calculations and analysis of packing and H-bonding

interactions support the helix assignment found in the cytochrome b6f structure

from green alga but not the assignment of TM helices in the cyanobacterium b6f

structure.

CONCLUSION: LIPS calculations can be used for the prediction of helix orientation

in ab initio modeling of polytopic membrane proteins. We also show with the

example of two cytochrome b6f structures that our method can identify

questionable helix assignments in membrane proteins. The LIPS server is available

online at http://gila.bioengr.uic.edu/lab/larisa/lips.html.

DOI: 10.1186/1472-6807-6-13

PMCID: PMC1540425

PMID: 16792816 [Indexed for MEDLINE]

2969. Bioinformation. 2006 Jun 15;1(5):156-7.

A web server for transcription factor binding site prediction.

Su G(1), Mao B, Wang J.

Author information:

(1)The State Key Laboratory of Pharmaceutical Biotechnology and Model Animal

Research Center, School of Life Science, Nanjing University, Nanjing - 210093,

China.

Promoter prediction has gained increased attention in studies related to

transcriptional regulation of gene expression. We developed a web server named

PMSearch (Poly Matrix Search) which utilizes Position Frequency Matrices (PFMs)

to predict transcription factor binding sites (TFBSs) in DNA sequences. PMSearch

takes PFMs (either user-defined or retrieved from local dataset which currently

contains 507 PFMs from Transfac Public 7.0 and JASPAR) and DNA sequences of

interest as the input, then scans the DNA sequences with PFMs and reports the

sites of high scores as the putative binding sites. The output of the server

includes 1) A plot for the distribution of predicted TFBS along the DNA sequence,

2) A table listing location, score and motif for each putative binding site, and

3) Clusters of predicted binding sites. PMSearch also provides links for

accessing clusters of PFMs that are similar to the input PFMs to facilitate

complicated promoter analysis.AVAILABILITY: PMSearch is available for free at

http://www.nicemice.cn/bioinfo/PMS.

PMCID: PMC1891680

PMID: 17597879

2970. BMC Bioinformatics. 2006 Jun 14;7:301.

Improving the accuracy of protein secondary structure prediction using structural

alignment.

Montgomerie S(1), Sundararaj S, Gallin WJ, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB, T6G 2E8,

Canada. montgomerie@shaw.ca

BACKGROUND: The accuracy of protein secondary structure prediction has steadily

improved over the past 30 years. Now many secondary structure prediction methods

routinely achieve an accuracy (Q3) of about 75%. We believe this accuracy could

be further improved by including structure (as opposed to sequence) database

comparisons as part of the prediction process. Indeed, given the large size of

the Protein Data Bank (>35,000 sequences), the probability of a newly identified

sequence having a structural homologue is actually quite high.

RESULTS: We have developed a method that performs structure-based sequence

alignments as part of the secondary structure prediction process. By mapping the

structure of a known homologue (sequence ID >25%) onto the query protein's

sequence, it is possible to predict at least a portion of that query protein's

secondary structure. By integrating this structural alignment approach with

conventional (sequence-based) secondary structure methods and then combining it

with a "jury-of-experts" system to generate a consensus result, it is possible to

attain very high prediction accuracy. Using a sequence-unique test set of 1644

proteins from EVA, this new method achieves an average Q3 score of 81.3%.

Extensive testing indicates this is approximately 4-5% better than any other

method currently available. Assessments using non sequence-unique test sets

(typical of those used in proteome annotation or structural genomics) indicate

that this new method can achieve a Q3 score approaching 88%.

CONCLUSION: By using both sequence and structure databases and by exploiting the

latest techniques in machine learning it is possible to routinely predict protein

secondary structure with an accuracy well above 80%. A program and web server,

called PROTEUS, that performs these secondary structure predictions is accessible

at http://wishart.biology.ualberta.ca/proteus. For high throughput or batch

sequence analyses, the PROTEUS programs, databases (and server) can be downloaded

and run locally.

DOI: 10.1186/1471-2105-7-301

PMCID: PMC1550433

PMID: 16774686 [Indexed for MEDLINE]

2971. BMC Struct Biol. 2006 Jun 7;6:11.

ProFace: a server for the analysis of the physicochemical features of

protein-protein interfaces.

Saha RP(1), Bahadur RP, Pal A, Mandal S, Chakrabarti P.

Author information:

(1)Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIIM, Calcutta

700 054, India. rudra@bic.boseinst.ernet.in

BACKGROUND: Molecular recognition is all pervasive in biology. Protein molecules

are involved in enzyme regulation, immune response, signal transduction, oligomer

assembly, etc. Delineation of physical and chemical features of the interface

formed by protein-protein association would allow us to better understand protein

interaction networks on one hand, and to design molecules that can engage a given

interface and thereby control protein function on the other hand.

RESULTS: ProFace is a suite of programs that uses a file, containing atomic

coordinates of a multi-chain molecule, as input and analyzes the interface

between any two or more subunits. The interface residues are shown segregated

into spatial patches (if such a clustering is possible based on an input

threshold distance) and/or core and rim regions. A number of physicochemical

parameters defining the interface is tabulated. Among the different output files,

one contains the list of interacting residues across the interface. Results can

be used to infer if a particular interface belongs to a homodimeric molecule.

CONCLUSION: A web-server, ProFace (available at

http://www.boseinst.ernet.in/resources/bioinfo/stag.html) has been developed for

dissecting protein-protein interfaces and deriving various physicochemical

parameters.

DOI: 10.1186/1472-6807-6-11

PMCID: PMC1513576

PMID: 16759379 [Indexed for MEDLINE]

2972. Comput Biol Chem. 2006 Jun;30(3):203-8.

Predicting O-glycosylation sites in mammalian proteins by using SVMs.

Li S(1), Liu B, Zeng R, Cai Y, Li Y.

Author information:

(1)Bioinformatics Center, Shanghai Institutes for Biological Sciences, Chinese

Academy of Sciences, China.

O-glycosylation is one of the most important, frequent and complex

post-translational modifications. This modification can activate and affect

protein functions. Here, we present three support vector machines models based on

physical properties, 0/1 system, and the system combining the above two features.

The prediction accuracies of the three models have reached 0.82, 0.85 and 0.85,

respectively. The accuracies of the three SVMs methods were evaluated by

'leave-one-out' cross validation. This approach provides a useful tool to help

identify the O-glycosylation sites in mammalian proteins. An online prediction

web server is available at http://www.biosino.org/Oglyc.

DOI: 10.1016/j.compbiolchem.2006.02.002

PMID: 16731044 [Indexed for MEDLINE]

2973. Proteins. 2006 Jun 1;63(4):868-77.

Scoring a diverse set of high-quality docked conformations: a metascore based on

electrostatic and desolvation interactions.

Camacho CJ(1), Ma H, Champ PC.

Author information:

(1)Department of Computational Biology, University of Pittsburgh, Pittsburgh,

Pennsylvania 15213, USA. ccamacho@pitt.edu

Predicting protein-protein interactions involves sampling and scoring docked

conformations. Barring some large structural rearrangement, rapidly sampling the

space of docked conformations is now a real possibility, and the limiting step

for the successful prediction of protein interactions is the scoring function

used to reduce the space of conformations from billions to a few, and eventually

one high affinity complex. An atomic level free-energy scoring function that

estimates in units of kcal/mol both electrostatic and desolvation interactions

(plus van der Waals if appropriate) of protein-protein docked conformations is

used to rerank the blind predictions (860 in total) submitted for six targets to

the community-wide Critical Assessment of PRediction of Interactions (CAPRI;

http://capri.ebi.ac.uk). We found that native-like models often have varying

intermolecular contacts and atom clashes, making unlikely that one can construct

a universal function that would rank all these models as native-like.

Nevertheless, our scoring function is able to consistently identify the

native-like complexes as those with the lowest free energy for the individual

models of 16 (out of 17) human predictors for five of the targets, while at the

same time the modelers failed to do so in more than half of the cases. The

scoring of high-quality models developed by a wide variety of methods and force

fields confirms that electrostatic and desolvation forces are the dominant

interactions determining the bound structure. The CAPRI experiment has shown that

modelers can predict valuable models of protein-protein complexes, and

improvements in scoring functions should soon solve the docking problem for

complexes whose backbones do not change much upon binding. A scoring server and

programs are available at http://structure.pitt.edu.

2006 Wiley-Liss, Inc.

DOI: 10.1002/prot.20932

PMID: 16506242 [Indexed for MEDLINE]

2974. Proteins. 2006 Jun 1;63(4):892-906.

On the nature of cavities on protein surfaces: application to the identification

of drug-binding sites.

Nayal M(1), Honig B.

Author information:

(1)Howard Hughes Medical Institute, Center for Computational Biology and

Bioinformatics, Department of Biochemistry and Molecular Biophysics, Columbia

University, New York, New York, USA.

In this article we introduce a new method for the identification and the accurate

characterization of protein surface cavities. The method is encoded in the

program SCREEN (Surface Cavity REcognition and EvaluatioN). As a first test of

the utility of our approach we used SCREEN to locate and analyze the surface

cavities of a nonredundant set of 99 proteins cocrystallized with drugs. We find

that this set of proteins has on average about 14 distinct cavities per protein.

In all cases, a drug is bound at one (and sometimes more than one) of these

cavities. Using cavity size alone as a criterion for predicting drug-binding

sites yields a high balanced error rate of 15.7%, with only 71.7% coverage. Here

we characterize each surface cavity by computing a comprehensive set of 408

physicochemical, structural, and geometric attributes. By applying modern machine

learning techniques (Random Forests) we were able to develop a classifier that

can identify drug-binding cavities with a balanced error rate of 7.2% and

coverage of 88.9%. Only 18 of the 408 cavity attributes had a statistically

significant role in the prediction. Of these 18 important attributes, almost all

involved size and shape rather than physicochemical properties of the surface

cavity. The implications of these results are discussed. A SCREEN Web server is

available at http://interface.bioc.columbia.edu/screen.

2006 Wiley-Liss, Inc.

DOI: 10.1002/prot.20897

PMID: 16477622 [Indexed for MEDLINE]

2975. BMC Genomics. 2006 May 31;7:132.

CMD: a Cotton Microsatellite Database resource for Gossypium genomics.

Blenda A(1), Scheffler J, Scheffler B, Palmer M, Lacape JM, Yu JZ, Jesudurai C,

Jung S, Muthukumar S, Yellambalase P, Ficklin S, Staton M, Eshelman R, Ulloa M,

Saha S, Burr B, Liu S, Zhang T, Fang D, Pepper A, Kumpatla S, Jacobs J, Tomkins

J, Cantrell R, Main D.

Author information:

(1)Department of Genetics and Biochemistry, Clemson University, Biosystems

Research Center, 51 New Cherry Street, Clemson, SC 29634, USA. blenda@clemson.edu

BACKGROUND: The Cotton Microsatellite Database (CMD) http://www.cottonssr.org is

a curated and integrated web-based relational database providing centralized

access to publicly available cotton microsatellites, an invaluable resource for

basic and applied research in cotton breeding.

DESCRIPTION: At present CMD contains publication, sequence, primer, mapping and

homology data for nine major cotton microsatellite projects, collectively

representing 5,484 microsatellites. In addition, CMD displays data for three of

the microsatellite projects that have been screened against a panel of core

germplasm. The standardized panel consists of 12 diverse genotypes including

genetic standards, mapping parents, BAC donors, subgenome representatives, unique

breeding lines, exotic introgression sources, and contemporary Upland cottons

with significant acreage. A suite of online microsatellite data mining tools are

accessible at CMD. These include an SSR server which identifies microsatellites,

primers, open reading frames, and GC-content of uploaded sequences; BLAST and

FASTA servers providing sequence similarity searches against the existing cotton

SSR sequences and primers, a CAP3 server to assemble EST sequences into longer

transcripts prior to mining for SSRs, and CMap, a viewer for comparing cotton SSR

maps.

CONCLUSION: The collection of publicly available cotton SSR markers in a

centralized, readily accessible and curated web-enabled database provides a more

efficient utilization of microsatellite resources and will help accelerate basic

and applied research in molecular breeding and genetic mapping in Gossypium spp.

DOI: 10.1186/1471-2164-7-132

PMCID: PMC1539020

PMID: 16737546 [Indexed for MEDLINE]

2976. BMC Bioinformatics. 2006 May 15;7:254.

Bounded search for de novo identification of degenerate cis-regulatory elements.

Carlson JM(1), Chakravarty A, Khetani RS, Gross RH.

Author information:

(1)Department of Computer Science and Engineering, University of Washington,

Seattle, WA 98105, USA. jcarlson@cs.washington.edu

BACKGROUND: The identification of statistically overrepresented sequences in the

upstream regions of coregulated genes should theoretically permit the

identification of potential cis-regulatory elements. However, in practice many

cis-regulatory elements are highly degenerate, precluding the use of an

exhaustive word-counting strategy for their identification. While numerous

methods exist for inferring base distributions using a position weight matrix,

recent studies suggest that the independence assumptions inherent in the model,

as well as the inability to reach a global optimum, limit this approach.

RESULTS: In this paper, we report PRISM, a degenerate motif finder that leverages

the relationship between the statistical significance of a set of binding sites

and that of the individual binding sites. PRISM first identifies overrepresented,

non-degenerate consensus motifs, then iteratively relaxes each one into a

high-scoring degenerate motif. This approach requires no tunable parameters,

thereby lending itself to unbiased performance comparisons. We therefore compare

PRISM's performance against nine popular motif finders on 28 well-characterized

S. cerevisiae regulons. PRISM consistently outperforms all other programs.

Finally, we use PRISM to predict the binding sites of uncharacterized regulons.

Our results support a proposed mechanism of action for the yeast cell-cycle

transcription factor Stb1, whose binding site has not been determined

experimentally.

CONCLUSION: The relationship between statistical measures of the binding sites

and the set as a whole leads to a simple means of identifying the diverse range

of cis-regulatory elements to which a protein binds. This approach leverages the

advantages of word-counting, in that position dependencies are implicitly

accounted for and local optima are more easily avoided. While we sacrifice

guaranteed optimality to prevent the exponential blowup of exhaustive search, we

prove that the error is bounded and experimentally show that the performance is

superior to other methods. A Java implementation of this algorithm can be

downloaded from our web server at http://genie.dartmouth.edu/prism.

DOI: 10.1186/1471-2105-7-254

PMCID: PMC1481619

PMID: 16700920 [Indexed for MEDLINE]

2977. Proteins. 2006 May 15;63(3):542-50.

Two-stage support vector regression approach for predicting accessible surface

areas of amino acids.

Nguyen MN(1), Rajapakse JC.

Author information:

(1)BioInformatics Research Centre, School of Computer Engineering, Nanyang

Technological University, Singapore.

We address the problem of predicting solvent accessible surface area (ASA) of

amino acid residues in protein sequences, without classifying them into buried

and exposed types. A two-stage support vector regression (SVR) approach is

proposed to predict real values of ASA from the position-specific scoring

matrices generated from PSI-BLAST profiles. By adding SVR as the second stage to

capture the influences on the ASA value of a residue by those of its neighbors,

the two-stage SVR approach achieves improvements of mean absolute errors up to

3.3%, and correlation coefficients of 0.66, 0.68, and 0.67 on the Manesh dataset

of 215 proteins, the Barton dataset of 502 nonhomologous proteins, and the Carugo

dataset of 338 proteins, respectively, which are better than the scores published

earlier on these datasets. A Web server for protein ASA prediction by using a

two-stage SVR method has been developed and is available

(http://birc.ntu.edu.sg/~ pas0186457/asa.html).

(c) 2006 Wiley-Liss, Inc.

DOI: 10.1002/prot.20883

PMID: 16456847 [Indexed for MEDLINE]

2978. J Chem Inf Model. 2006 May-Jun;46(3):971-84.

Computational Science and Engineering Online (CSE-Online): a cyber-infrastructure

for scientific computing.

Truong TN(1), Nayak M, Huynh HH, Cook T, Mahajan P, Tran LT, Bharath J, Jain S,

Pham HB, Boonyasiriwat C, Nguyen N, Andersen E, Kim Y, Choe S, Choi J, Cheatham

TE 3rd, Facelli JC.

Author information:

(1)Department of Chemistry and the Interdisciplinary Computational, Engineering,

and Science Program, Department of Medicinal Chemistry, University of Utah, 315

South 1400 East, Rm 2020, Salt Lake City, Utah 84112, USA. truong@chem.utah.edu

With the expansion of the Internet and World Wide Web (or the Web), research

environments have changed dramatically. As a result, the need to be able to

efficiently and securely access information and resources from remote computer

systems is becoming even more critical. This paper describes the development of

an extendable integrated Web-accessible simulation environment for computational

science and engineering called Computational Science and Engineering Online

(CSE-Online; http://cse-online.net). CSE-Online is based on a unique

client-server software architecture that can distribute the workload between the

client and server computers in such a way as to minimize the communication

between the client and server, thus making the environment less-sensitive to

network instability. Furthermore, the new software architecture allows the user

to access data and resources on one or more remote servers as well as on the

computing grid while having the full capability of the Web-services collaborative

environment. It can be accessed anytime and anywhere from a Web browser connected

to the network by either a wired or wireless connection. It has different modes

of operations to support different working environments and styles. CSE-Online is

evolving into middleware that can provide a framework for accessing and managing

remote data and resources including the computing grid for any domain, not

necessarily just within computational science and engineering.

DOI: 10.1021/ci0503917

PMID: 16711715 [Indexed for MEDLINE]

2979. Immunome Res. 2006 Apr 24;2:2.

Improved method for predicting linear B-cell epitopes.

Larsen JE(1), Lund O, Nielsen M.

Author information:

(1)Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208,

Technical University of Denmark, DK-2800 Kgs, Lyngby, Denmark. jepl@cbs.dtu.dk

BACKGROUND: B-cell epitopes are the sites of molecules that are recognized by

antibodies of the immune system. Knowledge of B-cell epitopes may be used in the

design of vaccines and diagnostics tests. It is therefore of interest to develop

improved methods for predicting B-cell epitopes. In this paper, we describe an

improved method for predicting linear B-cell epitopes.

RESULTS: In order to do this, three data sets of linear B-cell epitope annotated

proteins were constructed. A data set was collected from the literature, another

data set was extracted from the AntiJen database and a data sets of epitopes in

the proteins of HIV was collected from the Los Alamos HIV database. An unbiased

validation of the methods was made by testing on data sets on which they were

neither trained nor optimized on. We have measured the performance in a

non-parametric way by constructing ROC-curves.

CONCLUSION: The best single method for predicting linear B-cell epitopes is the

hidden Markov model. Combining the hidden Markov model with one of the best

propensity scale methods, we obtained the BepiPred method. When tested on the

validation data set this method performs significantly better than any of the

other methods tested. The server and data sets are publicly available at

http://www.cbs.dtu.dk/services/BepiPred.

DOI: 10.1186/1745-7580-2-2

PMCID: PMC1479323

PMID: 16635264

2980. Nihon Hoshasen Gijutsu Gakkai Zasshi. 2006 Apr 20;62(4):529-38.

[Design and development of a secure DICOM-Network Attached Server].

[Article in Japanese]

Tachibana H(1), Omatsu M, Higuchi K, Umeda T.

Author information:

(1)Radiology Department, Toranomon Hospital.

It is not easy to connect a Web-based server with an existing DICOM server, and

using a Web-based server on the Internet has risks. In this study, we designed

and developed a secure DICOM-Network Attached Server (DICOM-NAS) through which

the DICOM server in a hospital LAN was connected to the Internet. After receiving

a client's image export request, the DICOM-NAS sent it to the DICOM server using

the DICOM protocol. The server then provided DICOM images to the DICOM-NAS, which

transferred them to the client, using HTTP. The DICOM-NAS plays an important role

between the DICOM protocol and HTTP, and stores the requested images only

temporarily. The DICOM server keeps all of the original DICOM images. If an

unauthorized user attempts to access the DICOM-NAS, medical images cannot be

accessed because images are not stored in the DICOM-NAS. Furthermore, the

DICOM-NAS has features related to reporting and MPR. Therefore, the DICOM-NAS

does not require a large storage capacity, but can greatly improve information

security.

PMID: 16639395 [Indexed for MEDLINE]

2981. Bioinformatics. 2006 Apr 15;22(8):1024-6. Epub 2006 Feb 2.

The LCB Data Warehouse.

Ameur A(1), Yankovski V, Enroth S, Spjuth O, Komorowski J.

Author information:

(1)The Linnaeus Centre for Bioinformatics, Uppsala University, Sweden.

The Linnaeus Centre for Bioinformatics Data Warehouse (LCB-DWH) is a web-based

infrastructure for reliable and secure microarray gene expression data management

and analysis that provides an online service for the scientific community. The

LCB-DWH is an effort towards a complete system for storage (using the BASE

system), analysis and publication of microarray data. Important features of the

system include: access to established methods within R/Bioconductor for data

analysis, built-in connection to the Gene Ontology database and a scripting

facility for automatic recording and re-play of all the steps of the analysis.

The service is up and running on a high performance server. At present there are

more than 150 registered users.AVAILABILITY: An open functional version is

available at https://dw.lcb.uu.se/index.phtml?i\_login=test. User accounts are

created upon request. Additional facilities including plug-ins, user

documentation and a password protected data storage system are available from

http://www.lcb.uu.se/lcbdw.php

DOI: 10.1093/bioinformatics/btl036

PMID: 16455749 [Indexed for MEDLINE]

2982. Bioinformatics. 2006 Apr 15;22(8):924-33. Epub 2006 Jan 29.

Capturing expert knowledge with argumentation: a case study in bioinformatics.

Jefferys BR(1), Kelley LA, Sergot MJ, Fox J, Sternberg MJ.

Author information:

(1)Division of Molecular Biosciences, Imperial College, London SW7 2AZ, UK.

MOTIVATION: The output of a bioinformatic tool such as BLAST must usually be

interpreted by an expert before reliable conclusions can be drawn. This may be

based upon the expert's experience, additional data and statistical analysis.

Often the process is laborious, goes unrecorded and may be biased. Argumentation

is an established technique for reasoning about situations where absolute truth

or precise probability is impossible to determine.

RESULTS: We demonstrate the application of argumentation to 3D-PSSM, a protein

structure prediction tool. The expert's interpretation of results is represented

as an argumentation framework. Given a 3D-PSSM result, an automated procedure

constructs arguments for and against the conclusion that the result is a good

predictor of protein structure. In addition to capturing the unique expertise of

the author of 3D-PSSM for distribution to users, an improvement in recall of 5-10

percentage points is achieved. This technique can be applied to a wide range of

bioinformatic tools.

AVAILABILITY: Example public server and benchmarking data are available at

http://www.sbg.bio.ic.ac.uk/~brj03/argumentation/paper/. Source code available on

request.

DOI: 10.1093/bioinformatics/btl018

PMID: 16446279 [Indexed for MEDLINE]

2983. BMC Bioinformatics. 2006 Apr 3;7:185.

Ebbie: automated analysis and storage of small RNA cloning data using a dynamic

web server.

Ebhardt HA(1), Wiese KC, Unrau PJ.

Author information:

(1)Department of Molecular Biology and Biochemistry, Simon Fraser University,

8888 University Drive, V5A 1S6 Burnaby, B.C., Canada. aebhardt@sfu.ca

BACKGROUND: DNA sequencing is used ubiquitously: from deciphering genomes to

determining the primary sequence of small RNAs (smRNAs). The cloning of smRNAs is

currently the most conventional method to determine the actual sequence of these

important regulators of gene expression. Typical smRNA cloning projects involve

the sequencing of hundreds to thousands of smRNA clones that are delimited at

their 5' and 3' ends by fixed sequence regions. These primers result from the

biochemical protocol used to isolate and convert the smRNA into clonable PCR

products. Recently we completed a smRNA cloning project involving tobacco plants,

where analysis was required for approximately 700 smRNA sequences. Finding no

easily accessible research tool to enter and analyze smRNA sequences we developed

Ebbie to assist us with our study.

RESULTS: Ebbie is a semi-automated smRNA cloning data processing algorithm, which

initially searches for any substring within a DNA sequencing text file, which is

flanked by two constant strings. The substring, also termed smRNA or insert, is

stored in a MySQL and BlastN database. These inserts are then compared using

BlastN to locally installed databases allowing the rapid comparison of the insert

to both the growing smRNA database and to other static sequence databases. Our

laboratory used Ebbie to analyze scores of DNA sequencing data originating from

an smRNA cloning project. Through its built-in instant analysis of all inserts

using BlastN, we were able to quickly identify 33 groups of smRNAs from

approximately 700 database entries. This clustering allowed the easy

identification of novel and highly expressed clusters of smRNAs. Ebbie is

available under GNU GPL and currently implemented on

http://bioinformatics.org/ebbie/.

CONCLUSION: Ebbie was designed for medium sized smRNA cloning projects with about

1,000 database entries. Ebbie can be used for any type of sequence analysis where

two constant primer regions flank a sequence of interest. The reliable storage of

inserts, and their annotation in a MySQL database, BlastN comparison of new

inserts to dynamic and static databases make it a powerful new tool in any

laboratory using DNA sequencing. Ebbie also prevents manual mistakes during the

excision process and speeds up annotation and data-entry. Once the server is

installed locally, its access can be restricted to protect sensitive new DNA

sequencing data. Ebbie was primarily designed for smRNA cloning projects, but can

be applied to a variety of RNA and DNA cloning projects.

DOI: 10.1186/1471-2105-7-185

PMCID: PMC1450305

PMID: 16584563 [Indexed for MEDLINE]

2984. Bioinformatics. 2006 Apr 1;22(7):891-3. Epub 2006 Feb 2.

PROFbval: predict flexible and rigid residues in proteins.

Schlessinger A(1), Yachdav G, Rost B.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 650 West 168th Street BB217, New York, NY 10032, USA.

profbval@rostlab.org

The mobility of a residue on the protein surface is closely linked to its

function. The identification of extremely rigid or flexible surface residues can

therefore contribute information crucial for solving the complex problem of

identifying functionally important residues in proteins. Mobility is commonly

measured by B-value data from high-resolution three-dimensional X-ray structures.

Few methods predict B-values from sequence. Here, we present PROFbval, the first

web server to predict normalized B-values from amino acid sequence. The server

handles amino acid sequences (or alignments) as input and outputs normalized

B-value and two-state (flexible/rigid) predictions. The server also assigns a

reliability index for each prediction. For example, PROFbval correctly identifies

residues in active sites on the surface of enzymes as particularly

rigid.AVAILABILITY: http://www.rostlab.org/services/profbval

CONTACT: profbval@rostlab.org

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics

online.

DOI: 10.1093/bioinformatics/btl032

PMID: 16455751 [Indexed for MEDLINE]

2985. Bioinformatics. 2006 Apr 1;22(7):900-1. Epub 2006 Jan 19.

REMORA: a pilot in the ocean of BioMoby web-services.

Carrere S(1), Gouzy J.

Author information:

(1)INRA-CNRS Laboratoire des Interactions Plantes Micro-organismes (LIPM) BP

52627, 31326 Castanet Tolosan Cedex, France.

Emerging web-services technology allows interoperability between multiple

distributed architectures. Here, we present REMORA, a web server implemented

according to the BioMoby web-service specifications, providing life science

researchers with an easy-to-use workflow generator and launcher, a repository of

predefined workflows and a survey system.CONTACT: Jerome.Gouzy@toulouse.inra.fr

AVAILABILITY: The REMORA web server is freely available at

http://bioinfo.genopole-toulouse.prd.fr/remora, sources are available upon

request from the authors.

DOI: 10.1093/bioinformatics/btl001

PMID: 16423924 [Indexed for MEDLINE]

2986. Epigenetics. 2006 Apr-Jun;1(2):101-5. Epub 2006 Apr 5.

The MethDB DAS server: adding an epigenetic information layer to the human

genome.

Negre V(1), Grunau C.

Author information:

(1)Institut de Génétique Humaine, CNRS UPR 1142, Montpellier, France.

The DNA methylation database MethDB (www.methdb.net) was developed in order to

standardize and collect the dispersed data about this epigenetic phenomenon in a

common resource. In the first version of MethDB, data was gathered by annotators

and the database could only be queried. In a second step, we added an on-line

data submission system that is open to the public. Here we present the DAS

annotation server of MethDB that allows integration of MethDB into the network of

biological databases via the Distributed Annotation System (DAS) and the

representation of DNA methylation data as an epigenetic information layer to the

human genome. In order to validate our system and to incorporate the data of the

first large scale methylation analysis of the human genome, we assembled the

31312 sequences of the human CpG island tagging project into 13786 CpG islands

and imported them into MethDB. The database contains now 19905 methylation

content data and 5382 methylation patterns or profiles for 48 species, 1511

individuals, 198 tissues and cell lines and 79 phenotypes.

PMID: 17965614 [Indexed for MEDLINE]

2987. IEEE/ACM Trans Comput Biol Bioinform. 2006 Apr-Jun;3(2):98-113.

Bayesian segmental models with multiple sequence alignment profiles for protein

secondary structure and contact map prediction.

Chu W(1), Ghahramani Z, Podtelezhnikov A, Wild DL.

Author information:

(1)Gatsby Computational Neuroscience Unit, University College London, London, UK.

chuwei@gatsby.ucl.ac.uk

In this paper, we develop a segmental semi-Markov model (SSMM) for protein

secondary structure prediction which incorporates multiple sequence alignment

profiles with the purpose of improving the predictive performance. The segmental

model is a generalization of the hidden Markov model where a hidden state

generates segments of various length and secondary structure type. A novel

parameterized model is proposed for the likelihood function that explicitly

represents multiple sequence alignment profiles to capture the segmental

conformation. Numerical results on benchmark data sets show that incorporating

the profiles results in substantial improvements and the generalization

performance is promising. By incorporating the information from long range

interactions in beta-sheets, this model is also capable of carrying out inference

on contact maps. This is an important advantage of probabilistic generative

models over the traditional discriminative approach to protein secondary

structure prediction. The Web server of our algorithm and supplementary materials

are available at http://public.kgi.edu/-wild/bsm.html.

DOI: 10.1109/TCBB.2006.17

PMID: 17048397 [Indexed for MEDLINE]

2988. J Bioinform Comput Biol. 2006 Apr;4(2):589-96.

RNAKinetics: a web server that models secondary structure kinetics of an

elongating RNA.

Danilova LV(1), Pervouchine DD, Favorov AV, Mironov AA.

Author information:

(1)Institute for Problems of Information Transition RAS, Bolshoi Karetnyi per.

19, Moscow, 127994, Russia. dlv2k@mail.ru

The RNAKinetics server (http://www.ig-msk.ru/RNA/kinetics) is a web interface for

the newly developed RNAKinetics software. The software models the dynamics of RNA

secondary structure by the means of kinetic analysis of folding transitions of a

growing RNA molecule. The result of the modeling is a kinetic ensemble, i.e. a

collection of RNA structures that are endowed with probabilities, which depend on

time. This approach gives comprehensive probabilistic description of RNA folding

pathways, revealing important kinetic details that are not captured by the

traditional structure prediction methods. The access to the RNAKinetics server is

free.

PMID: 16819804 [Indexed for MEDLINE]

2989. J Lipid Res. 2006 Apr;47(4):824-31. Epub 2006 Jan 27.

Prediction of the functional class of lipid binding proteins from

sequence-derived properties irrespective of sequence similarity.

Lin HH(1), Han LY, Zhang HL, Zheng CJ, Xie B, Chen YZ.

Author information:

(1)Bioinformatics and Drug Design Group, Department of Computational Science,

National University of Singapore, Singapore 117543.

Lipid binding proteins play important roles in signaling, regulation, membrane

trafficking, immune response, lipid metabolism, and transport. Because of their

functional and sequence diversity, it is desirable to explore additional methods

for predicting lipid binding proteins irrespective of sequence similarity. This

work explores the use of support vector machines (SVMs) as such a method. SVM

prediction systems are developed using 14,776 lipid binding and 133,441 nonlipid

binding proteins and are evaluated by an independent set of 6,768 lipid binding

and 64,761 nonlipid binding proteins. The computed prediction accuracy is 78.9,

79.5, 82.2, 79.5, 84.4, 76.6, 90.6, 79.0, and 89.9% for lipid degradation, lipid

metabolism, lipid synthesis, lipid transport, lipid binding, lipopolysaccharide

biosynthesis, lipoprotein, lipoyl, and all lipid binding proteins, respectively.

The accuracy for the nonmember proteins of each class is 99.9, 99.2, 99.6, 99.8,

99.9, 99.8, 98.5, 99.9, and 97.0%, respectively. Comparable accuracies are

obtained when homologous proteins are considered as one, or by using a different

SVM kernel function. Our method predicts 86.8% of the 76 lipid binding proteins

nonhomologous to any protein in the Swiss-Prot database and 89.0% of the 73 known

lipid binding domains as lipid binding. These findings suggest the usefulness of

SVMs for facilitating the prediction of lipid binding proteins. Our software can

be accessed at the SVMProt server

(http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi).

DOI: 10.1194/jlr.M500530-JLR200

PMID: 16443826 [Indexed for MEDLINE]

2990. Proteins. 2006 Apr 1;63(1):1-5.

Prediction of buried helices in multispan alpha helical membrane proteins.

Adamian L(1), Liang J.

Author information:

(1)Department of Bioengineering, University of Illinois, Chicago, Illinois

60612-7340, USA.

Analysis of a database of structures of membrane proteins shows that membrane

proteins composed of 10 or more transmembrane (TM) helices often contain buried

helices that are inaccessible to phospholipids. We introduce a method for

identifying TM helices that are least phospholipid accessible and for prediction

of fully buried TM helices in membrane proteins from sequence information alone.

Our method is based on the calculation of residue lipophilicity and evolutionary

conservation. Given that the number of buried helices in a membrane protein is

known, our method achieves an accuracy of 78% and a Matthew's correlation

coefficient of 0.68. A server for this tool (RANTS) is available online at

http://gila.bioengr.uic.edu/lab/.

2006 Wiley-Liss, Inc.

DOI: 10.1002/prot.20874

PMID: 16419070 [Indexed for MEDLINE]

2991. Psychiatr Prax. 2006 Apr;33(3):117-23.

[Online-consulting for eating disorders--analysis of users and contents].

[Article in German]

Grunwald M(1), Wesemann D.

Author information:

(1)Universität Leipzig, Klinik für Psychiatrie. mgrun@medizin.uni-leipzig.de

OBJECTIVE: Since 1998, the online information and consulting server for patients

with eating disorders and their relatives (www.ab-server.de) offers an online

consulting service.

METHODS: 2176 e-mails were qualitatively and quantitatively analysed.

RESULTS: The symptom descriptions refer mostly to bulimia nervosa (63.1%). People

mainly asked for behaviour patterns in dealing with the illness or with an

affected person (33.3%) as well as for information about the illness (18.7%).

CONCLUSIONS: The low threshold and professional online consulting service is

highly accepted by the target group.

DOI: 10.1055/s-2005-915244

PMID: 16583346 [Indexed for MEDLINE]

2992. Bioinformatics. 2006 Mar 15;22(6):773-4. Epub 2006 Jan 19.

SUSPECTS: enabling fast and effective prioritization of positional candidates.

Adie EA(1), Adams RR, Evans KL, Porteous DJ, Pickard BS.

Author information:

(1)Medical Genetics Section, School of Molecular and Clinical Medicine,

University of Edinburgh EH4 2XU UK. euan.adie@ed.ac.uk

SUSPECTS is a web-based server which combines annotation and sequence-based

approaches to prioritize disease candidate genes in large regions of interest. It

uses multiple lines of evidence to rank genes quickly and effectively while

limiting the effect of annotation bias to significantly improve

performance.AVAILABILITY: SUSPECTS is freely available at

http://www.genetics.med.ed.ac.uk/suspects/

SUPPLEMENTARY INFORMATION: A quick-start guide in Macromedia Flash format is

available at http://www.genetics.med.ed.ac.uk/suspects/help.shtml and Excel

spreadsheets detailing the comparative performance of the software are included

as Supplementary material.

DOI: 10.1093/bioinformatics/btk031

PMID: 16423925 [Indexed for MEDLINE]

2993. Bioinformatics. 2006 Mar 15;22(6):762-4. Epub 2006 Jan 10.

GUUGle: a utility for fast exact matching under RNA complementary rules including

G-U base pairing.

Gerlach W(1), Giegerich R.

Author information:

(1)Faculty of Technology, Bielefeld University 33615 Bielefeld, Germany.

MOTIVATION: RNA secondary structure analysis often requires searching for

potential helices in large sequence data.

RESULTS: We present a utility program GUUGle that efficiently locates potential

helical regions under RNA base pairing rules, which include Watson-Crick as well

as G-U pairs. It accepts a positive and a negative set of sequences, and

determines all exact matches under RNA rules between positive and negative

sequences that exceed a specified length. The GUUGle algorithm can also be

adapted to use a precomputed suffix array of the positive sequence set. We show

how this program can be effectively used as a filter preceding a more

computationally expensive task such as miRNA target prediction.

AVAILABILITY: GUUGle is available via the Bielefeld Bioinformatics Server at

http://bibiserv.techfak.uni-bielefeld.de/guugle

DOI: 10.1093/bioinformatics/btk041

PMID: 16403789 [Indexed for MEDLINE]

2994. Proteins. 2006 Mar 15;62(3):617-29.

Large-scale prediction of disulphide bridges using kernel methods,

two-dimensional recursive neural networks, and weighted graph matching.

Cheng J(1), Saigo H, Baldi P.

Author information:

(1)Institute for Genomics and Bioinformatics, School of Information and Computer

Sciences, University of California, Irvine 92697, USA.

The formation of disulphide bridges between cysteines plays an important role in

protein folding, structure, function, and evolution. Here, we develop new methods

for predicting disulphide bridges in proteins. We first build a large curated

data set of proteins containing disulphide bridges to extract relevant

statistics. We then use kernel methods to predict whether a given protein chain

contains intrachain disulphide bridges or not, and recursive neural networks to

predict the bonding probabilities of each pair of cysteines in the chain. These

probabilities in turn lead to an accurate estimation of the total number of

disulphide bridges and to a weighted graph matching problem that can be addressed

efficiently to infer the global disulphide bridge connectivity pattern. This

approach can be applied both in situations where the bonded state of each

cysteine is known, or in ab initio mode where the state is unknown. Furthermore,

it can easily cope with chains containing an arbitrary number of disulphide

bridges, overcoming one of the major limitations of previous approaches. It can

classify individual cysteine residues as bonded or nonbonded with 87% specificity

and 89% sensitivity. The estimate for the total number of bridges in each chain

is correct 71% of the times, and within one from the true value over 94% of the

times. The prediction of the overall disulphide connectivity pattern is exact in

about 51% of the chains. In addition to using profiles in the input to leverage

evolutionary information, including true (but not predicted) secondary structure

and solvent accessibility information yields small but noticeable improvements.

Finally, once the system is trained, predictions can be computed rapidly on a

proteomic or protein-engineering scale. The disulphide bridge prediction server

(DIpro), software, and datasets are available through

www.igb.uci.edu/servers/psss.html.

(c) 2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20787

PMID: 16320312 [Indexed for MEDLINE]

2995. BMC Bioinformatics. 2006 Mar 2;7:104.

SCOWLP: a web-based database for detailed characterization and visualization of

protein interfaces.

Teyra J(1), Doms A, Schroeder M, Pisabarro MT.

Author information:

(1)Department of Bioinformatics, BIOTEC TU Dresden, Tatzberg 47-51, 01307

Dresden, Germany. joan.teyra@biotec.tu-dresden.de

BACKGROUND: Currently there is a strong need for methods that help to obtain an

accurate description of protein interfaces in order to be able to understand the

principles that govern molecular recognition and protein function. Many of the

recent efforts to computationally identify and characterize protein networks

extract protein interaction information at atomic resolution from the PDB.

However, they pay none or little attention to small protein ligands and solvent.

They are key components and mediators of protein interactions and fundamental for

a complete description of protein interfaces. Interactome profiling requires the

development of computational tools to extract and analyze protein-protein,

protein-ligand and detailed solvent interaction information from the PDB in an

automatic and comparative fashion. Adding this information to the existing one on

protein-protein interactions will allow us to better understand protein

interaction networks and protein function.

DESCRIPTION: SCOWLP (Structural Characterization Of Water, Ligands and Proteins)

is a user-friendly and publicly accessible web-based relational database for

detailed characterization and visualization of the PDB protein interfaces. The

SCOWLP database includes proteins, peptidic-ligands and interface water molecules

as descriptors of protein interfaces. It contains currently 74,907 protein

interfaces and 2,093,976 residue-residue interactions formed by 60,664 structural

units (protein domains and peptidic-ligands) and their interacting solvent. The

SCOWLP web-server allows detailed structural analysis and comparisons of protein

interfaces at atomic level by text query of PDB codes and/or by navigating a

SCOP-based tree. It includes a visualization tool to interactively display the

interfaces and label interacting residues and interface solvent by atomic

physicochemical properties. SCOWLP is automatically updated with every SCOP

release.

CONCLUSION: SCOWLP enriches substantially the description of protein interfaces

by adding detailed interface information of peptidic-ligands and solvent to the

existing protein-protein interaction databases. SCOWLP may be of interest to many

structural bioinformaticians. It provides a platform for automatic global mapping

of protein interfaces at atomic level, representing a useful tool for

classification of protein interfaces, protein binding comparative studies,

reconstruction of protein complexes and understanding protein networks. The

web-server with the database and its additional summary tables used for our

analysis are available at http://www.scowlp.org.

DOI: 10.1186/1471-2105-7-104

PMCID: PMC1459204

PMID: 16512892 [Indexed for MEDLINE]

2996. Acta Crystallogr A. 2006 Mar;62(Pt 2):115-28. Epub 2006 Feb 18.

Bilbao Crystallographic Server. II. Representations of crystallographic point

groups and space groups.

Aroyo MI(1), Kirov A, Capillas C, Perez-Mato JM, Wondratschek H.

Author information:

(1)Departamento de Física de la Materia Condensada, Universidad del País Vasco,

E-48080 Bilbao, Spain.

The Bilbao Crystallographic Server is a web site with crystallographic programs

and databases freely available on-line (http://www.cryst.ehu.es). The server

gives access to general information related to crystallographic symmetry groups

(generators, general and special positions, maximal subgroups, Brillouin zones

etc.). Apart from the simple tools for retrieving the stored data, there are

programs for the analysis of group-subgroup relations between space groups

(subgroups and supergroups, Wyckoff-position splitting schemes etc.). There are

also software packages studying specific problems of solid-state physics,

structural chemistry and crystallography. This article reports on the programs

treating representations of point and space groups. There are tools for the

construction of irreducible representations, for the study of the correlations

between representations of group-subgroup pairs of space groups and for the

decompositions of Kronecker products of representations.

DOI: 10.1107/S0108767305040286

PMID: 16489249

2997. Comput Methods Programs Biomed. 2006 Mar;81(3):197-202. Epub 2006 Feb 28.

Design and development of a secure DICOM-Network Attached Server.

Tachibana H(1), Omatsu M, Higuchi K, Umeda T.

Author information:

(1)Medical Image Engineering, Kitasato University, Graduate School of Medical

Sciences, 1-15-1 Kitasato, Sagamihara, Kanagawa 228-8555, Japan.

tatibana@umeken3.ahs.kitasato-u.ac.jp

It is not easy to connect a web-based server with an existing DICOM server, and

using a web-based server on the INTERNET has risks. In this study, we designed

and developed the secure DICOM-Network Attached Server (DICOM-NAS) through which

the DICOM server in a hospital-Local Area Network (LAN) was connected to the

INTERNET. After receiving a Client's image export request, the DICOM-NAS sent it

to the DICOM server with DICOM protocol. The server then provided DICOM images to

the DICOM-NAS, which transferred them to the Client using HTTP. The DICOM-NAS

plays an important role between DICOM protocol and HTTP, and only temporarily

stores the requested images. The DICOM server keeps all of the original DICOM

images. When unwanted outsiders attempt to get into the DICOM-NAS, they cannot

access any medical images because these images are not stored in the DICOM-NAS.

Therefore, the DICOM-NAS does not require large storage, but can greatly improve

information security.

DOI: 10.1016/j.cmpb.2005.11.015

PMID: 16503366 [Indexed for MEDLINE]

2998. Plant Physiol. 2006 Mar;140(3):818-29.

AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription

factors into regulatory networks.

Palaniswamy SK(1), James S, Sun H, Lamb RS, Davuluri RV, Grotewold E.

Author information:

(1)Human Cancer Genetics Program, Comprehensive Cancer Center, Department of

Molecular Virology, Immunology and Medical Genetics , The Ohio State University,

Columbus, 43210, USA.

Gene regulatory pathways converge at the level of transcription, where

interactions among regulatory genes and between regulators and target genes

result in the establishment of spatiotemporal patterns of gene expression. The

growing identification of direct target genes for key transcription factors (TFs)

through traditional and high-throughput experimental approaches has facilitated

the elucidation of regulatory networks at the genome level. To integrate this

information into a Web-based knowledgebase, we have developed the Arabidopsis

Gene Regulatory Information Server (AGRIS). AGRIS, which contains all Arabidopsis

(Arabidopsis thaliana) promoter sequences, TFs, and their target genes and

functions, provides the scientific community with a platform to establish

regulatory networks. AGRIS currently houses three linked databases: AtcisDB

(Arabidopsis thaliana cis-regulatory database), AtTFDB (Arabidopsis thaliana

transcription factor database), and AtRegNet (Arabidopsis thaliana regulatory

network). AtTFDB contains 1,690 Arabidopsis TFs and their sequences (protein and

DNA) grouped into 50 (October 2005) families with information on available

mutants in the corresponding genes. AtcisDB consists of 25,806 (September 2005)

promoter sequences of annotated Arabidopsis genes with a description of putative

cis-regulatory elements. AtRegNet links, in direct interactions, several hundred

genes with the TFs that control their expression. The current release of AtRegNet

contains a total of 187 (September 2005) direct targets for 66 TFs. AGRIS can be

accessed at http://Arabidopsis.med.ohio-state.edu.

DOI: 10.1104/pp.105.072280

PMCID: PMC1400579

PMID: 16524982 [Indexed for MEDLINE]

2999. Proteins. 2006 Mar 1;62(4):881-91.

SSALN: an alignment algorithm using structure-dependent substitution matrices and

gap penalties learned from structurally aligned protein pairs.

Qiu J(1), Elber R.

Author information:

(1)Department of Computer Science, Cornell University, Ithaca, New York 14853,

USA.

In template-based modeling of protein structures, the generation of the alignment

between the target and the template is a critical step that significantly affects

the accuracy of the final model. This paper proposes an alignment algorithm SSALN

that learns substitution matrices and position-specific gap penalties from a

database of structurally aligned protein pairs. In addition to the amino acid

sequence information, secondary structure and solvent accessibility information

of a position are used to derive substitution scores and position-specific gap

penalties. In a test set of CASP5 targets, SSALN outperforms sequence alignment

methods such as a Smith-Waterman algorithm with BLOSUM50 and PSI\_BLAST. SSALN

also generates better alignments than PSI\_BLAST in the CASP6 test set. LOOPP

server prediction based on an SSALN alignment is ranked the best for target

T0280\_1 in CASP6. SSALN is also compared with several threading methods and

sequence alignment methods on the ProSup benchmark. SSALN has the highest

alignment accuracy among the methods compared. On the Fischer's benchmark, SSALN

performs better than CLUSTALW and GenTHREADER, and generates more alignments with

accuracy >50%, >60% or >70% than FUGUE, but fewer alignments with accuracy >80%

than FUGUE. All the supplemental materials can be found at

http://www.cs.cornell.edu/ approximately jianq/research.htm.

2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20854

PMID: 16385554 [Indexed for MEDLINE]

3000. Proteins. 2006 Mar 1;62(4):1125-32.

Prediction of protein stability changes for single-site mutations using support

vector machines.

Cheng J(1), Randall A, Baldi P.

Author information:

(1)Institute for Genomics and Bioinformatics, School of Information and Computer

Sciences, University of California, Irvine, California 92697-3425, USA.

Accurate prediction of protein stability changes resulting from single amino acid

mutations is important for understanding protein structures and designing new

proteins. We use support vector machines to predict protein stability changes for

single amino acid mutations leveraging both sequence and structural information.

We evaluate our approach using cross-validation methods on a large dataset of

single amino acid mutations. When only the sign of the stability changes is

considered, the predictive method achieves 84% accuracy-a significant improvement

over previously published results. Moreover, the experimental results show that

the prediction accuracy obtained using sequence alone is close to the accuracy

obtained using tertiary structure information. Because our method can accurately

predict protein stability changes using primary sequence information only, it is

applicable to many situations where the tertiary structure is unknown, overcoming

a major limitation of previous methods which require tertiary information. The

web server for predictions of protein stability changes upon mutations (MUpro),

software, and datasets are available at

http://www.igb.uci.edu/servers/servers.html.

2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20810

PMID: 16372356 [Indexed for MEDLINE]

3001. Tuberculosis (Edinb). 2006 Mar;86(2):115-24. Epub 2005 Jul 21.

ProPred analysis and experimental evaluation of promiscuous T-cell epitopes of

three major secreted antigens of Mycobacterium tuberculosis.

Mustafa AS(1), Shaban FA.

Author information:

(1)Department of Microbiology, Faculty of Medicine, Kuwait University, PO Box

24923, Safat 13110, Kuwait. abusalim@hsc.edu.kw

In the search for safe vaccine candidates against tuberculosis (TB), subunit

vaccines including peptide-based candidates deserve consideration. However, an

important requirement for such vaccine candidates is their promiscuous

presentation to Th1 cells mediating protective immunity against TB, i.e. Th1

cells secreting IFN-gamma. The aim of the present study was to identify

promiscuous Th1 cell epitopes of three major secreted antigens of Mycobacterium

tuberculosis, i.e. ESAT-6, CFP10 and MPT70 by using a virtual matrix-based

prediction program (ProPred) for peptide binding to 51 HLA-DR alleles. The

ProPred analysis of these proteins was performed using the server

(http:www.imtech.res.in/raghava/ProPed/). The peptides predicted to bind > 50%

HLA-DR alleles included in the ProPred were considered promiscuous for binding

predictions. Based on this criteria, one region in ESAT-6 (aa 69-77), two regions

in CFP10 (aa 55-66 and aa 76-84) and four regions in MPT70 (aa 1-11, aa 81-95, aa

124-140 and aa 182-191) were considered promiscuous HLA-DR binders. The

experimental evaluation of these regions, by using overlapping synthetic peptides

for presentation to T-cells, confirmed the promiscuous nature of peptides

covering the regions aa 69-77, aa 76-84 and aa 182-191 of ESAT-6, CFP10 and

MPT70, respectively. These results demonstrate that the ProPred analysis can

facilitate the selection of promiscuous peptides recognized by Th1 cells, and

thus it can be useful in the identification of peptide-based vaccine candidates

against TB.

DOI: 10.1016/j.tube.2005.05.001

PMID: 16039905 [Indexed for MEDLINE]

3002. Virchows Arch. 2006 Mar;448(3):248-55. Epub 2005 Nov 22.

Factors to keep in mind when introducing virtual microscopy.

Glatz-Krieger K(1), Spornitz U, Spatz A, Mihatsch MJ, Glatz D.

Author information:

(1)Institute for Pathology, University Hospital of Basel, Basel, Switzerland.

Katharina.Glatz@unibas.ch

Digitization of glass slides and delivery of so-called virtual slides (VS)

emulating a real microscope over the Internet have become reality due to recent

improvements in technology. We have implemented a virtual microscope for

instruction of medical students and for continuing medical education. Up to

30,000 images per slide are captured using a microscope with an automated stage.

The images are post-processed and then served by a plain hypertext transfer

protocol (http)-server. A virtual slide client (vMic) based on Macromedia's Flash

MX, a highly accepted technology available on every modern Web browser, has been

developed. All necessary virtual slide parameters are stored in an XML file

together with the image. Evaluation of the courses by questionnaire indicated

that most students and many but not all pathologists regard virtual slides as an

adequate replacement for traditional slides. All our virtual slides are publicly

accessible over the World Wide Web (WWW) at http://vmic.unibas.ch . Recently,

several commercially available virtual slide acquisition systems (VSAS) have been

developed that use various technologies to acquire and distribute virtual slides.

These systems differ in speed, image quality, compatibility, viewer

functionalities and price. This paper gives an overview of the factors to keep in

mind when introducing virtual microscopy.

DOI: 10.1007/s00428-005-0112-2

PMID: 16362822 [Indexed for MEDLINE]

3003. BMC Bioinformatics. 2006 Feb 17;7:79.

Taxonomic colouring of phylogenetic trees of protein sequences.

Palidwor G(1), Reynaud EG, Andrade-Navarro MA.

Author information:

(1)Ontario Genomics Innovation Centre, Ottawa Health Research Institute, 501

Smyth, Ottawa, ON K1H 8L6, Canada. gpalidwor@ohri.ca

BACKGROUND: Phylogenetic analyses of protein families are used to define the

evolutionary relationships between homologous proteins. The interpretation of

protein-sequence phylogenetic trees requires the examination of the taxonomic

properties of the species associated to those sequences. However, there is no

online tool to facilitate this interpretation, for example, by automatically

attaching taxonomic information to the nodes of a tree, or by interactively

colouring the branches of a tree according to any combination of taxonomic

divisions. This is especially problematic if the tree contains on the order of

hundreds of sequences, which, given the accelerated increase in the size of the

protein sequence databases, is a situation that is becoming common.

RESULTS: We have developed PhyloView, a web based tool for colouring phylogenetic

trees upon arbitrary taxonomic properties of the species represented in a protein

sequence phylogenetic tree. Provided that the tree contains SwissProt, SpTrembl,

or GenBank protein identifiers, the tool retrieves the taxonomic information from

the corresponding database. A colour picker displays a summary of the findings

and allows the user to associate colours to the leaves of the tree according to

any number of taxonomic partitions. Then, the colours are propagated to the

branches of the tree.

CONCLUSION: PhyloView can be used at http://www.ogic.ca/projects/phyloview/. A

tutorial, the software with documentation, and GPL licensed source code, can be

accessed at the same web address.

DOI: 10.1186/1471-2105-7-79

PMCID: PMC1386715

PMID: 16503967 [Indexed for MEDLINE]

3004. BMC Bioinformatics. 2006 Feb 16;7:75.

Detecting overlapping coding sequences in virus genomes.

Firth AE(1), Brown CM.

Author information:

(1)Department of Biochemistry, University of Otago, PO Box 56, Dunedin, New

Zealand. aef@sanger.otago.ac.nz

BACKGROUND: Detecting new coding sequences (CDSs) in viral genomes can be

difficult for several reasons. The typically compact genomes often contain a

number of overlapping coding and non-coding functional elements, which can result

in unusual patterns of codon usage; conservation between related sequences can be

difficult to interpret--especially within overlapping genes; and viruses often

employ non-canonical translational mechanisms--e.g. frameshifting, stop codon

read-through, leaky-scanning and internal ribosome entry sites--which can conceal

potentially coding open reading frames (ORFs).

RESULTS: In a previous paper we introduced a new statistic--MLOGD (Maximum

Likelihood Overlapping Gene Detector)--for detecting and analysing overlapping

CDSs. Here we present (a) an improved MLOGD statistic, (b) a greatly extended

suite of software using MLOGD, (c) a database of results for 640 virus sequence

alignments, and (d) a web-interface to the software and database. Tests show

that, from an alignment with just 20 mutations, MLOGD can discriminate

non-overlapping CDSs from non-coding ORFs with a typical accuracy of up to 98%,

and can detect CDSs overlapping known CDSs with a typical accuracy of 90%. In

addition, the software produces a variety of statistics and graphics, useful for

analysing an input multiple sequence alignment.

CONCLUSION: MLOGD is an easy-to-use tool for virus genome annotation, detecting

new CDSs--in particular overlapping or short CDSs--and for analysing overlapping

CDSs following frameshift sites. The software, web-server, database and

supplementary material are available at http://guinevere.otago.ac.nz/mlogd.html.

DOI: 10.1186/1471-2105-7-75

PMCID: PMC1395342

PMID: 16483358 [Indexed for MEDLINE]

3005. Bioinformatics. 2006 Feb 1;22(3):303-9. Epub 2005 Nov 17.

Enhanced recognition of protein transmembrane domains with prediction-based

structural profiles.

Cao B(1), Porollo A, Adamczak R, Jarrell M, Meller J.

Author information:

(1)Biomedical Informatics, Children's Hospital Research Foundation, Cincinnati,

OH 45229, USA.

MOTIVATION: Membrane domain prediction has recently been re-evaluated by several

groups, suggesting that the accuracy of existing methods is still rather limited.

In this work, we revisit this problem and propose novel methods for prediction of

alpha-helical as well as beta-sheet transmembrane (TM) domains. The new approach

is based on a compact representation of an amino acid residue and its

environment, which consists of predicted solvent accessibility and secondary

structure of each amino acid. A recently introduced method for solvent

accessibility prediction trained on a set of soluble proteins is used here to

indicate segments of residues that are predicted not to be accessible to water

and, therefore, may be 'buried' in the membrane. While evolutionary profiles in

the form of a multiple alignment are used to derive these simple 'structural

profiles', they are not used explicitly for the membrane domain prediction and

the overall number of parameters in the model is significantly reduced. This

offers the possibility of a more reliable estimation of the free parameters in

the model with a limited number of experimentally resolved membrane protein

structures.

RESULTS: Using cross-validated training on available sets of structurally

resolved and non-redundant alpha and beta membrane proteins, we demonstrate that

membrane domain prediction methods based on such a compact representation

outperform approaches that utilize explicitly evolutionary profiles and multiple

alignments. Moreover, using an external evaluation by the TMH Benchmark server we

show that our final prediction protocol for the TM helix prediction is

competitive with the state-of-the-art methods, achieving per-residue accuracy of

approximately 89% and per-segment accuracy of approximately 80% on the set of

high resolution structures used by the TMH Benchmark server. At the same time the

observed rates of confusion with signal peptides and globular proteins are the

lowest among the tested methods. The new method is available online at

http://minnou.cchmc.org.

DOI: 10.1093/bioinformatics/bti784

PMID: 16293670 [Indexed for MEDLINE]

3006. Genomics Proteomics Bioinformatics. 2006 Feb;4(1):42-7.

VICMpred: an SVM-based method for the prediction of functional proteins of

Gram-negative bacteria using amino acid patterns and composition.

Saha S(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Chandigarh, India.

In this study, an attempt has been made to predict the major functions of

gram-negative bacterial proteins from their amino acid sequences. The dataset

used for training and testing consists of 670 non-redundant gram-negative

bacterial proteins (255 of cellular process, 60 of information molecules, 285 of

metabolism, and 70 of virulence factors). First we developed an SVM-based method

using amino acid and dipeptide composition and achieved the overall accuracy of

52.39% and 47.01%, respectively. We introduced a new concept for the

classification of proteins based on tetrapeptides, in which we identified the

unique tetrapeptides significantly found in a class of proteins. These

tetrapeptides were used as the input feature for predicting the function of a

protein and achieved the overall accuracy of 68.66%. We also developed a hybrid

method in which the tetrapeptide information was used with amino acid composition

and achieved the overall accuracy of 70.75%. A five-fold cross validation was

used to evaluate the performance of these methods. The web server VICMpred has

been developed for predicting the function of gram-negative bacterial proteins

(http://www.imtech.res.in/raghava/vicmpred/).

DOI: 10.1016/S1672-0229(06)60015-6

PMCID: PMC5054027

PMID: 16689701 [Indexed for MEDLINE]

3007. Proteins. 2006 Feb 1;62(2):343-55.

Will my protein crystallize? A sequence-based predictor.

Smialowski P(1), Schmidt T, Cox J, Kirschner A, Frishman D.

Author information:

(1)Department of Genome Oriented Bioinformatics, Technische Universität München,

Wissenschaftszentrum Weihenstephan, Freising, Germany.

We propose a machine-learning approach to sequence-based prediction of protein

crystallizability in which we exploit subtle differences between proteins whose

structures were solved by X-ray analysis [or by both X-ray and nuclear magnetic

resonance (NMR) spectroscopy] and those proteins whose structures were solved by

NMR spectroscopy alone. Because the NMR technique is usually applied on

relatively small proteins, sequence length distributions of the X-ray and NMR

datasets were adjusted to avoid predictions biased by protein size. As feature

space for classification, we used frequencies of mono-, di-, and tripeptides

represented by the original 20-letter amino acid alphabet as well as by several

reduced alphabets in which amino acids were grouped by their physicochemical and

structural properties. The classification algorithm was constructed as a

two-layered structure in which the output of primary support vector machine

classifiers operating on peptide frequencies was combined by a second-level Naive

Bayes classifier. Due to the application of metamethods for cost sensitivity, our

method is able to handle real datasets with unbalanced class representation. An

overall prediction accuracy of 67% [65% on the positive (crystallizable) and 69%

on the negative (noncrystallizable) class] was achieved in a 10-fold

cross-validation experiment, indicating that the proposed algorithm may be a

valuable tool for more efficient target selection in structural genomics. A Web

server for protein crystallizability prediction called SECRET is available at

http://webclu.bio.wzw.tum.de:8080/secret.

2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20789

PMID: 16315316 [Indexed for MEDLINE]

3008. BMC Bioinformatics. 2006 Jan 19;7:27.

NOXclass: prediction of protein-protein interaction types.

Zhu H(1), Domingues FS, Sommer I, Lengauer T.

Author information:

(1)Max-Planck-lnstitut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken,

Germany. hzhu@mpi-sb.mpg.de

BACKGROUND: Structural models determined by X-ray crystallography play a central

role in understanding protein-protein interactions at the molecular level.

Interpretation of these models requires the distinction between non-specific

crystal packing contacts and biologically relevant interactions. This has been

investigated previously and classification approaches have been proposed.

However, less attention has been devoted to distinguishing different types of

biological interactions. These interactions are classified as obligate and

non-obligate according to the effect of the complex formation on the stability of

the protomers. So far no automatic classification methods for distinguishing

obligate, non-obligate and crystal packing interactions have been made available.

RESULTS: Six interface properties have been investigated on a dataset of 243

protein interactions. The six properties have been combined using a support

vector machine algorithm, resulting in NOXclass, a classifier for distinguishing

obligate, non-obligate and crystal packing interactions. We achieve an accuracy

of 91.8% for the classification of these three types of interactions using a

leave-one-out cross-validation procedure.

CONCLUSION: NOXclass allows the interpretation and analysis of protein quaternary

structures. In particular, it generates testable hypotheses regarding the nature

of protein-protein interactions, when experimental results are not available. We

expect this server will benefit the users of protein structural models, as well

as protein crystallographers and NMR spectroscopists. A web server based on the

method and the datasets used in this study are available at

http://noxclass.bioinf.mpi-inf.mpg.de/.

DOI: 10.1186/1471-2105-7-27

PMCID: PMC1386716

PMID: 16423290 [Indexed for MEDLINE]

3009. Bioinformatics. 2006 Jan 15;22(2):129-33. Epub 2005 Nov 14.

'Protein Peeling': an approach for splitting a 3D protein structure into compact

fragments.

Gelly JC(1), de Brevern AG, Hazout S.

Author information:

(1)INSERM U726, Equipe de Bioinformatique Génomique and Moléculaire (EBGM),

Université Denis Diderot-Paris 7, case 7113, 75251 Paris Cedex 05, France.

MOTIVATION: The object of this study is to propose a new method to identify small

compact units that compose protein three-dimensional structures. These fragments,

called 'protein units (PU)', are a new level of description to well understand

and analyze the organization of protein structures. The method only works from

the contact probability matrix, i.e. the inter Calpha-distances translated into

probabilities. It uses the principle of conventional hierarchical clustering,

leading to a series of nested partitions of the 3D structure. Every step aims at

dividing optimally a unit into 2 or 3 subunits according to a criterion called

'partition index' assessing the structural independence of the subunits newly

defined. Moreover, an entropy-derived squared correlation R is used for assessing

globally the protein structure dissection. The method is compared to other

splitting algorithms and shows relevant performance.

AVAILABILITY: An Internet server with dedicated tools is available at

http://www.ebgm.jussieu.fr/~gelly/

DOI: 10.1093/bioinformatics/bti773

PMID: 16301202 [Indexed for MEDLINE]

3010. Bioinformatics. 2006 Jan 15;22(2):188-94. Epub 2005 Nov 8.

Residue-rotamer-reduction algorithm for the protein side-chain conformation

problem.

Xie W(1), Sahinidis NV.

Author information:

(1)Department of Chemical and Biomolecular Engineering, University of Illinois at

Urbana-Champaign 600 South Mathews Avenue, Urbana, IL 61801, USA.

MOTIVATION: The protein side-chain conformation problem is a central problem in

proteomics with wide applications in protein structure prediction and design.

Computational complexity results show that the problem is hard to solve. Yet,

instances from realistic applications are large and demand fast and reliable

algorithms.

RESULTS: We propose a new global optimization algorithm, which for the first time

integrates residue reduction and rotamer reduction techniques previously

developed for the protein side-chain conformation problem. We show that the

proposed approach simplifies dramatically the topology of the underlining residue

graph. Computations show that our algorithm solves problems using only 1-10% of

the time required by the mixed-integer linear programming approach available in

the literature. In addition, on a set of hard side-chain conformation problems,

our algorithm runs 2-78 times faster than SCWRL 3.0, which is widely used for

solving these problems.

AVAILABILITY: The implementation is available as an online server at

http://eudoxus.scs.uiuc.edu/r3.html

DOI: 10.1093/bioinformatics/bti763

PMID: 16278239 [Indexed for MEDLINE]

3011. Bioinformatics. 2006 Jan 15;22(2):181-7. Epub 2005 Nov 2.

SSEP-Domain: protein domain prediction by alignment of secondary structure

elements and profiles.

Gewehr JE(1), Zimmer R.

Author information:

(1)Practical Informatics and Bioinformatics Group, Department of Informatics,

Ludwig-Maximilians-University Amalienstrasse 17, D-80333 Munich, Germany.

jan.gewehr@bio.ifi.lmu.de

MOTIVATION: The prediction of protein domains is a crucial task for functional

classification, homology-based structure prediction and structural genomics. In

this paper, we present the SSEP-Domain protein domain prediction approach, which

is based on the application of secondary structure element alignment (SSEA) and

profile-profile alignment (PPA) in combination with InterPro pattern searches.

SSEA allows rapid screening for potential domain regions while PPA provides us

with the necessary specificity for selecting significant hits. The combination

with InterPro patterns allows finding domain regions without solved structural

templates if sequence family definitions exist.

RESULTS: A preliminary version of SSEP-Domain was ranked among the top-performing

domain prediction servers in the CASP 6 and CAFASP 4 experiments. Evaluation of

the final version shows further improvement over these results together with a

significant speed-up.

AVAILABILITY: The server is available at http://www.bio.ifi.lmu.de/SSEP/

DOI: 10.1093/bioinformatics/bti751

PMID: 16267083 [Indexed for MEDLINE]

3012. AMIA Annu Symp Proc. 2006:851.

UMLSKS SUGGEST: an auto-complete feature for the UMLSKS interface using AJAX.

Bangalore A(1), Browne A, Divita G.

Author information:

(1)US National Library of Medicine, Bethesda, Maryland, USA.

The idea behind any auto-complete feature is to suggest words or phrases which

are likely to complete what the user is typing. The user can then select a

suggestion and avoid typing it in full. This feature by all accounts has enhanced

the usability of web interfaces. Typically browsers implement this feature by

caching a fixed number of queries, previously entered by the user on the client

side. Google Suggest (http://ajaxpatterns.org/Suggestion) is an application that

replaces a browser's auto-complete feature with one specific to Google searching.

Implementing a google suggest like feature requires a large server

infrastructure, as basically every time you type a letter, a database is being

hit. One of the ways that this feature can be implemented by many applications

without access to a large server infrastructure would be to create a small list

of suggestions that contains the most likely suggestions for any given

application In this poster we describe the methodology used to create a list of

suggestions for the UMLS Knowledge Source Server (UMLSKS) interface. We believe

that this methodology can be used by any web based application to create a list

of suggestions.

PMCID: PMC1839508

PMID: 17238471 [Indexed for MEDLINE]

3013. Appl Bioinformatics. 2006;5(4):225-36.

Ontology annotation treebrowser : an interactive tool where the complementarity

of medical subject headings and gene ontology improves the interpretation of gene

lists.

Bresell A(1), Servenius B, Persson B.

Author information:

(1)IFM Bioinformatics, Linköping University, Linköping, Sweden.

andersb@ifm.liu.se

Gene expression and proteomics analysis allow the investigation of thousands of

biomolecules in parallel. This results in a long list of interesting genes or

proteins and a list of annotation terms in the order of thousands. It is not a

trivial task to understand such a gene list and it would require extensive

efforts to bring together the overwhelming amounts of associated information from

the literature and databases. Thus, it is evident that we need ways of condensing

and filtering this information. An excellent way to represent knowledge is to use

ontologies, where it is possible to group genes or terms with overlapping

context, rather than studying one-dimensional lists of keywords. Therefore, we

have built the ontology annotation treebrowser (OAT) to represent, condense,

filter and summarise the knowledge associated with a list of genes or proteins.

The OAT system consists of two disjointed parts; a MySQL database named OATdb,

and a treebrowser engine that is implemented as a web interface. The OAT system

is implemented using Perl scripts on an Apache web server and the gene, ontology

and annotation data is stored in a relational MySQL database. In OAT, we have

harmonized the two ontologies of medical subject headings (MeSH) and gene

ontology (GO), to enable us to use knowledge both from the literature and the

annotation projects in the same tool. OAT includes multiple gene identifier sets,

which are merged internally in the OAT database. We have also generated novel

MeSH annotations by mapping accession numbers to MEDLINE entries. The ontology

browser OAT was created to facilitate the analysis of gene lists. It can be

browsed dynamically, so that a scientist can interact with the data and govern

the outcome. Test statistics show which branches are enriched. We also show that

the two ontologies complement each other, with surprisingly low overlap, by

mapping annotations to the Unified Medical Language System. We have developed a

novel interactive annotation browser that is the first to incorporate both MeSH

and GO for improved interpretation of gene lists. With OAT, we illustrate the

benefits of combining MeSH and GO for understanding gene lists. OAT is available

as a public web service at: http://www.ifm.liu.se/bioinfo/oat.

PMID: 17140269 [Indexed for MEDLINE]

3014. Appl Bioinformatics. 2006;5(3):193-8.

REGANOR: a gene prediction server for prokaryotic genomes and a database of high

quality gene predictions for prokaryotes.

Linke B(1), McHardy AC, Neuweger H, Krause L, Meyer F.

Author information:

(1)CeBiTec, Universität Bielefeld, Bielefeld, Germany.

With >1,000 prokaryotic genome sequencing projects ongoing or already finished,

comprehensive comparative analysis of the gene content of these genomes has

become viable. To allow for a meaningful comparative analysis, gene prediction of

the various genomes should be as accurate as possible. It is clear that improving

the state of genome annotation requires automated gene identification methods to

cope with the influence of artifacts, such as genomic GC content. There is

currently still room for improvement in the state of annotations. We present a

web server and a database of high-quality gene predictions. The web server is a

resource for gene identification in prokaryote genome sequences. It implements

our previously described, accurate gene finding method REGANOR. We also provide

novel gene predictions for 241 complete, or almost complete, prokaryotic genomes.

We demonstrate how this resource can easily be utilised to identify promising

candidates for currently missing genes from genome annotations with several

examples. All data sets are available online.AVAILABILITY: The gene finding

server is accessible via

https://www.cebitec.uni-bielefeld.de/groups/brf/software/reganor/cgi-bin/reganor\_

upload.cgi. The server software is available with the GenDB genome annotation

system (version 2.2.1 onwards) under the GNU general public license. The software

can be downloaded from https://sourceforge.net/projects/gendb/. More information

on installing GenDB and REGANOR and the system requirements can be found on the

GenDB project page

http://www.cebitec.uni-bielefeld.de/groups/brf/software/wiki/GenDBWiki/Administra

torDocumentation/GenDBInstallation

PMID: 16922601 [Indexed for MEDLINE]

3015. Appl Bioinformatics. 2006;5(2):125-30.

WebaCGH: an interactive online tool for the analysis and display of array

comparative genomic hybridisation data.

Frankenberger C(1), Wu X, Harmon J, Church D, Gangi LM, Munroe DJ, Urzúa U.

Author information:

(1)Laboratory of Molecular Technology, Scientific Application International

Corporation (SAIC) - Frederick, Inc., National Cancer Institute - Frederick,

Frederick, Maryland, USA.

Gene copy number variations occur both in normal cells and in numerous

pathologies including cancer and developmental diseases. Array comparative

genomic hybridisation (aCGH) is an emerging technology that allows detection of

chromosomal gains and losses in a high-resolution format. When aCGH is performed

on cDNA and oligonucleotide microarrays, the impact of DNA copy number on gene

transcription profiles may be directly compared. We have created an online

software tool, WebaCGH, that functions to (i) upload aCGH and gene transcription

results from multiple experiments; (ii) identify significant aberrant regions

using a local Z-score threshold in user-selected chromosomal segments subjected

to smoothing with moving averages; and (iii) display results in a graphical

format with full genome and individual chromosome views. In the individual

chromosome display, data can be zoomed in/out in both dimensions (i.e. ratio and

physical location) and plotted features can have 'mouse over' linking to outside

databases to identify loci of interest. Uploaded data can be stored indefinitely

for subsequent retrieval and analysis. WebaCGH was created as a Java-based web

application using the open-source database MySQL.AVAILABILITY: WebaCGH is freely

accessible at http://129.43.22.27/WebaCGH/welcome.htm

CONTACT: Xiaolin Wu (forestwu@mail.nih.gov) or Ulises Urzúa

(uurzua@med.uchile.cl).

PMID: 16722779 [Indexed for MEDLINE]

3016. Appl Bioinformatics. 2006;5(2):121-4.

OligoMatcher: analysis and selection of specific oligonucleotide sequences for

gene silencing by antisense or siRNA.

Mamidipalli S(1), Palakal M, Li S.

Author information:

(1)Department of Computer and Information Science, Indiana University - Purdue

University Indianapolis, Indianapolis, Indiana, USA.

OligoMatcher is a web-based tool for analysis and selection of unique

oligonucleotide sequences for gene silencing by antisense oligonucleotides (ASOs)

or small interfering RNA (siRNA). A specific BLAST server was built for analysing

sequences of ASOs that target pre-mRNA in the cell nucleus. Tissue- and

cell-specific expression data of potential cross-reactive genes are integrated in

the OligoMatcher program, which allows biologists to select unique

oligonucleotide sequences for their target genes in specific experimental

systems.AVAILABILITY: The OligoMatcher web server is available at

http://shelob.cs.iupui.edu:18081/oligomatch.php. The source code is freely

available for non-profit use on request to the authors.

CONTACT: Mathew Palakal (mpalakal@cs.iupui.edu) or Shuyu Li

(li\_shuyu\_dan@lilly.com).

PMID: 16722778 [Indexed for MEDLINE]

3017. Appl Bioinformatics. 2006;5(1):63-6.

Biotool2Web: creating simple Web interfaces for bioinformatics applications.

Shahid M(1), Alam I, Fuellen G.

Author information:

(1)International NRW Graduate School in Bioinformatics and Genome Research,

Center of Biotechnology (CeBiTec), University of Bielefeld, Bielefeld, Germany.

Currently there are many bioinformatics applications being developed, but there

is no easy way to publish them on the World Wide Web. We have developed a Perl

script, called Biotool2Web, which makes the task of creating web interfaces for

simple ('home-made') bioinformatics applications quick and easy. Biotool2Web uses

an XML document containing the parameters to run the tool on the Web, and

generates the corresponding HTML and common gateway interface (CGI) files ready

to be published on a web server.AVAILABILITY: This tool is available for download

at URL http://www.uni-muenster.de/Bioinformatics/services/biotool2web/

CONTACT: Georg Fuellen (fuellen@alum.mit.edu).

PMID: 16539540 [Indexed for MEDLINE]

3018. Appl Bioinformatics. 2006;5(1):55-61.

MHCPred 2.0: an updated quantitative T-cell epitope prediction server.

Guan P(1), Hattotuwagama CK, Doytchinova IA, Flower DR.

Author information:

(1)Edward Jenner Institute for Vaccine Research, Compton, Berkshire, UK.

The accurate computational prediction of T-cell epitopes can greatly reduce the

experimental overhead implicit in candidate epitope identification within genomic

sequences. In this article we present MHCPred 2.0, an enhanced version of our

online, quantitative T-cell epitope prediction server. The previous version of

MHCPred included mostly alleles from the human leukocyte antigen A (HLA-A) locus.

In MHCPred 2.0, mouse models are added and computational constraints removed.

Currently the server includes 11 human HLA class I, three human HLA class II, and

three mouse class I models. Additionally, a binding model for the human

transporter associated with antigen processing (TAP) is incorporated into the new

MHCPred. A tool for the design of heteroclitic peptides is also included within

the server. To refine the veracity of binding affinities prediction, a confidence

percentage is also now calculated for each peptide predicted.AVAILABILITY: As

previously, MHCPred 2.0 is freely available at the URL

http://www.jenner.ac.uk/MHCPred/

CONTACT: Darren R. Flower (darren.flower@jenner.ac.uk).

PMID: 16539539 [Indexed for MEDLINE]

3019. Appl Bioinformatics. 2006;5(1):45-7.

AMarge: Automated Extensive Quality Assessment of Affymetrix chips.

Lozano JJ(1), Kalko SG.

Author information:

(1)Laboratori de Microarrays, Centre de Regulació Genomica (CRG), Barcelona,

Spain. juanjo.lozano@crg.es

AMarge is a web tool for the automatic quality assessment of Affymetrix GeneChip

data. It is essential to have a trustworthy set of chips in order to derive gene

expression data for phenotypic analysis, and AMarge provides a complete and

rigorous web-accessible tool to fulfill this need. The quality assessment steps

include image plots of weights derived from a robust linear model fit of the

data, a 3'/5' RNA digestion plot, and Affymetrix Microarray Suite version 5.0

(MAS 5.0) quality standard procedures. Furthermore, robust multi-array average

expression values are generated in order to have a start-up expression set for

the subsequent analysis. The results of the complete analysis are summarised and

returned as an HTML report.AVAILABILITY: The AMarge web interface is accessible

at http://nin.crg.es/cgi-binf/AMargeWeb.cgi. A mirror server is also available at

http://bioinformatics.istge.it/AMarge-bin/AMargeWeb.cgi. The software

implementing all these methods is part of the Bioconductor project

(http://www.bioconductor.org).

PMID: 16539537 [Indexed for MEDLINE]

3020. In Silico Biol. 2006;6(4):307-10.

MACO: a gapped-alignment scoring tool for comparing transcription factor binding

sites.

Su G(1), Mao B, Wang J.

Author information:

(1)The State Key Laboratory of Pharmaceutical Biotechnology and Model Animal

Institute, Department of Biochemistry, Nanjing University, Nanjing 210093, PR

China.

We have implemented a novel gapped-alignment algorithm to compare Position

Frequency Matrices (PFMs) for Transcription Factor Binding Sites. The application

compares an input PFM with those collected from public databases and outputs

similarity scores, sequence alignments and related PFM clusters. MACO is freely

accessible on a web server located at www.nicemice.cn/bioinfo/MACO. Source code

is distributed upon request to the authors.

PMID: 16922693 [Indexed for MEDLINE]

3021. In Silico Biol. 2006;6(1-2):111-25.

Prediction of C alpha-H...O and C alpha-H...pi interactions in proteins using

recurrent neural network.

Kaur H(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector 39A, Chandigarh, India.

In this study, an attempt has been made to develop a method for predicting weak

hydrogen bonding interactions, namely, C alpha-H...O and C alpha-H...pi

interactions in proteins using artificial neural network. Both standard

feed-forward neural network (FNN) and recurrent neural networks (RNN) have been

trained and tested using five-fold cross-validation on a non-homologous dataset

of 2298 protein chains where no pair of sequences has more than 25% sequence

identity. It has been found that the prediction accuracy varies with the

separation distance between donor and acceptor residues. The maximum sensitivity

achieved with RNN for C alpha-H...O is 51.2% when donor and acceptor residues are

four residues apart (i.e. at delta D-A = 4) and for C alpha-H...pi is 82.1% at

delta D-A = 3. The performance of RNN is increased by 1-3% for both types of

interactions when PSIPRED predicted protein secondary structure is used. Overall,

RNN performs better than feed-forward networks at all separation distances

between donor-acceptor pair for both types of interactions. Based on the

observations, a web server CHpredict (available at

http://www.imtech.res.in/raghava/chpredict/) has been developed for predicting

donor and acceptor residues in C alpha-H...O and C alpha-H...pi interactions in

proteins.

PMID: 16789918 [Indexed for MEDLINE]

3022. J Mol Graph Model. 2006 Jan;24(4):296-306. Epub 2005 Nov 11.

A connection rule for alpha-carbon coarse-grained elastic network models using

chemical bond information.

Jeong JI(1), Jang Y, Kim MK.

Author information:

(1)Department of Mechanical Engineering, The Johns Hopkins University, Baltimore,

MD 21218, USA.

A sparser but more efficient connection rule (called a bond-cutoff method) for a

simplified alpha-carbon coarse-grained elastic network model is presented. One of

conventional connection rules for elastic network models is the distance-cutoff

method, where virtual springs connect an alpha-carbon with all neighbor

alpha-carbons within predefined distance-cutoff value. However, though the

maximum interaction distance between alpha-carbons is reported as 7 angstroms,

this cutoff value can make the elastic network unstable in many cases of protein

structures. Thus, a larger cutoff value (>11 angstroms) is often used to

establish a stable elastic network model in previous researches. To overcome this

problem, a connection rule for backbone model is proposed, which satisfies the

minimum condition to stabilize an elastic network. Based on the backbone

connections, each type of chemical interactions is considered and added to the

elastic network model: disulfide bonds, hydrogen bonds, and salt-bridges. In

addition, the van der Waals forces between alpha-carbons are modeled by using the

distance-cutoff method. With the proposed connection rule, one can make an

elastic network model with less than 7 angstroms distance cutoff, which can

reveal protein flexibility more sharply. Moreover, the normal modes from the new

elastic network model can reflect conformational changes of a given protein

better than ones by the distance-cutoff method. This method can save the

computational cost when calculating normal modes of a given protein structure,

because it can reduce the total number of connections. As a validation, six

example proteins are tested. Computational times and the overlap values between

the conformational change and infinitesimal motion calculated by normal mode

analysis are presented. Those animations are also available at UMass Morph Server

(http://biomechanics.ecs.umass.edu/umms.html).

DOI: 10.1016/j.jmgm.2005.09.006

PMID: 16289973 [Indexed for MEDLINE]

3023. Nucleic Acids Res. 2006;34(19):5660-9. Epub 2006 Oct 11.

Rational design and rapid screening of antisense oligonucleotides for prokaryotic

gene modulation.

Shao Y(1), Wu Y, Chan CY, McDonough K, Ding Y.

Author information:

(1)Wadsworth Center, New York State Department of Health, 150 New Scotland

Avenue, Albany, NY 12208, USA.

Antisense oligodeoxynucleotides (oligos) are widely used for functional studies

of both prokaryotic and eukaryotic genes. However, the identification of

effective target sites is a major issue in antisense applications. Here, we study

a number of thermodynamic and structural parameters that may affect the potency

of antisense inhibition. We develop a cell-free assay for rapid oligo screening.

This assay is used for measuring the expression of Escherichia coli lacZ, the

antisense target for experimental testing and validation. Based on a training set

of 18 oligos, we found that structural accessibility predicted by local folding

of the target mRNA is the most important predictor for antisense activity. This

finding was further confirmed by a direct validation study. In this study, a set

of 10 oligos was designed to target accessible sites, and another set of 10

oligos was selected to target inaccessible sites. Seven of the 10 oligos for

accessible sites were found to be effective (>50% inhibition), but none of the

oligos for inaccessible sites was effective. The difference in the antisense

activity between the two sets of oligos was statistically significant. We also

found that the predictability of antisense activity by target accessibility was

greatly improved for oligos targeted to the regions upstream of the end of the

active domain for beta-galactosidase, the protein encoded by lacZ. The

combination of the structure-based antisense design and extension of the lacZ

assay to include gene fusions will be applicable to high-throughput gene

functional screening, and to the identification of new drug targets in pathogenic

microbes. Design tools are available through the Sfold Web server at

http://sfold.wadsworth.org.

DOI: 10.1093/nar/gkl715

PMCID: PMC1636493

PMID: 17038332 [Indexed for MEDLINE]

3024. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D90-4.

HTPSELEX--a database of high-throughput SELEX libraries for transcription factor

binding sites.

Jagannathan V(1), Roulet E, Delorenzi M, Bucher P.

Author information:

(1)Swiss Institute for Experimental Cancer Research and Swiss Institute of

Bioinformatics, Ch. des Boveresses 155, CH-1066 Epalinges, Switzerland.

HTPSELEX is a public database providing access to primary and derived data from

high-throughput SELEX experiments aimed at characterizing the binding specificity

of transcription factors. The resource is primarily intended to serve

computational biologists interested in building models of transcription factor

binding sites from large sets of binding sequences. The guiding principle is to

make available all information that is relevant for this purpose. For each

experiment, we try to provide accurate information about the protein material

used, details of the wet lab protocol, an archive of sequencing trace files,

assembled clone sequences (concatemers) and complete sets of in vitro selected

protein-binding tags. In addition, we offer in-house derived binding sites

models. HTPSELEX also offers reasonably large SELEX libraries obtained with

conventional low-throughput protocols. The FTP site contains the trace archives

and database flatfiles. The web server offers user-friendly interfaces for

viewing individual entries and quality-controlled download of SELEX sequence

libraries according to a user-defined sequencing quality threshold. HTPSELEX is

available from ftp://ftp.isrec.isb-sib.ch/pub/databases/htpselex/ and

http://www.isrec.isb-sib.ch/htpselex.

DOI: 10.1093/nar/gkj049

PMCID: PMC1347412

PMID: 16381982 [Indexed for MEDLINE]

3025. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D82-5.

EPD in its twentieth year: towards complete promoter coverage of selected model

organisms.

Schmid CD(1), Perier R, Praz V, Bucher P.

Author information:

(1)Swiss Institute of Bioinformatics, Chemin des Boveresses 155, CH-1066

Epalinges, Switzerland.

The Eukaryotic Promoter Database (EPD) is an annotated non-redundant collection

of eukaryotic POL II promoters, experimentally defined by a transcription start

site (TSS). Access to promoter sequences is provided by pointers to positions in

the corresponding genomes. Promoter evidence comes from conventional TSS mapping

experiments for individual genes, or, starting from release 73, from mass genome

annotation projects. Subsets of promoter sequences with customized 5' and 3'

extensions can be downloaded from the EPD website. The focus of current

development efforts is to reach complete promoter coverage for important model

organisms as soon as possible. To speed up this process, a new class of

preliminary promoter entries has been introduced as of release 83, which requires

less stringent admission criteria. As part of a continuous integration process,

new web-based interfaces have been developed, which allow joint analysis of

promoter sequences with other bioinformatics resources developed by our group, in

particular programs offered by the Signal Search Analysis Server, and gene

expression data stored in the CleanEx database. EPD can be accessed at

http://www.epd.isb-sib.ch.

DOI: 10.1093/nar/gkj146

PMCID: PMC1347508

PMID: 16381980 [Indexed for MEDLINE]

3026. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D452-5.

Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene

Order Browser.

Byrne KP(1), Wolfe KH.

Author information:

(1)Department of Genetics, Smurfit Institute, University of Dublin, Trinity

College, Dublin 2, Ireland. kevin.byrne@tcd.ie

The Yeast Gene Order Browser (YGOB) is an online tool designed to facilitate the

comparative genomic visualization and appraisal of synteny within and between the

genomes of seven hemiascomycete yeast species. Three of these genomes are

polyploid, and hence contain intra-genomic syntenic regions, the correct assembly

of which is a particular success of YGOB. Designed to accurately assemble,

display and score gene order relationships, YGOB is both an interactive tool for

browsing genomic data, and a software engine now being used for evolutionary

analyses on a whole-genome scale. Underlying the online interface is the YGOB

database, which consists of homology assignments across the species, extensively

curated based on sequence similarity and novelly, an appraisal of genomic context

(synteny) in multiple genomes. Currently the YGOB database incorporates genome

data from Saccharomyces cerevisiae, Candida glabrata, Saccharomyces castellii,

Ashbya gossypii, Kluyveromyces lactis, Kluyveromyces waltii and Saccharomyces

kluyveri, but the system is scaleable to accommodate additional genomes. This

paper discusses the usage and utility of version 1.0 of YGOB, which is publicly

available at http://wolfe.gen.tcd.ie/ygob.

DOI: 10.1093/nar/gkj041

PMCID: PMC1347404

PMID: 16381909 [Indexed for MEDLINE]

3027. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D349-53.

INVHOGEN: a database of homologous invertebrate genes.

Paulsen I(1), von Haeseler A.

Author information:

(1)Department of Bioinformatics, Institute for Computer Sciences,

Heinrich-Heine-University Duesseldorf, Universitaetsstrasse 1, 40225 Duesseldorf,

Germany. paulseni@uni-duesseldorf.de

Classification of proteins into families of homologous sequences constitutes the

basis of functional analysis or of evolutionary studies. Here we present

INVertebrate HOmologous GENes (INVHOGEN), a database combining the available

invertebrate protein genes from UniProt (consisting of Swiss-Prot and TrEMBL)

into gene families. For each family INVHOGEN provides a multiple protein

alignment, a maximum likelihood based phylogenetic tree and taxonomic information

about the sequences. It is possible to download the corresponding GenBank

flatfiles, the alignment and the tree in Newick format. Sequences and related

information have been structured in an ACNUC database under a client/server

architecture. Thus, complex selections can be performed. An external graphical

tool (FamFetch) allows access to the data to evaluate homology relationships

between genes and distinguish orthologous from paralogous sequences. Thus,

INVHOGEN complements the well-known HOVERGEN database. The databank is available

at http://www.bi.uni-duesseldorf.de/~invhogen/invhogen.html.

DOI: 10.1093/nar/gkj100

PMCID: PMC1347462

PMID: 16381884 [Indexed for MEDLINE]

3028. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D296-301.

The Database of Macromolecular Motions: new features added at the decade mark.

Flores S(1), Echols N, Milburn D, Hespenheide B, Keating K, Lu J, Wells S, Yu EZ,

Thorpe M, Gerstein M.

Author information:

(1)Department of Physics, Yale University, P.O. Box 208120, New Haven, CT

06520-8120, USA. Samuel.Flores@yale.edu

The database of molecular motions, MolMovDB (http://molmovdb.org), has been in

existence for the past decade. It classifies macromolecular motions and provides

tools to interpolate between two conformations (the Morph Server) and predict

possible motions in a single structure. In 2005, we expanded the services offered

on MolMovDB. In particular, we further developed the Morph Server to produce

improved interpolations between two submitted structures. We added support for

multiple chains to the original adiabatic mapping interpolation, allowing the

analysis of subunit motions. We also added the option of using FRODA

interpolation, which allows for more complex pathways, potentially overcoming

steric barriers. We added an interface to a hinge prediction service, which acts

on single structures and predicts likely residue points for flexibility. We

developed tools to relate such points of flexibility in a structure to particular

key residue positions, i.e. active sites or highly conserved positions. Lastly,

we began relating our motion classification scheme to function using descriptions

from the Gene Ontology Consortium.

DOI: 10.1093/nar/gkj046

PMCID: PMC1347409

PMID: 16381870 [Indexed for MEDLINE]

3029. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D291-5.

MODBASE: a database of annotated comparative protein structure models and

associated resources.

Pieper U(1), Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom

M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A.

Author information:

(1)Department of Biopharmaceutical Sciences, California Institute for

Quantitative Biomedical Research, QB3 at Mission Bay, Office 503B, University of

California at San Francisco 1700 4th Street, San Francisco, CA 94158, USA.

MODBASE (http://salilab.org/modbase) is a database of annotated comparative

protein structure models for all available protein sequences that can be matched

to at least one known protein structure. The models are calculated by MODPIPE, an

automated modeling pipeline that relies on MODELLER for fold assignment,

sequence-structure alignment, model building and model assessment

(http:/salilab.org/modeller). MODBASE is updated regularly to reflect the growth

in protein sequence and structure databases, and improvements in the software for

calculating the models. MODBASE currently contains 3 094 524 reliable models for

domains in 1 094 750 out of 1 817 889 unique protein sequences in the UniProt

database (July 5, 2005); only models based on statistically significant

alignments and models assessed to have the correct fold despite insignificant

alignments are included. MODBASE also allows users to generate comparative models

for proteins of interest with the automated modeling server MODWEB

(http://salilab.org/modweb). Our other resources integrated with MODBASE include

comprehensive databases of multiple protein structure alignments (DBAli,

http://salilab.org/dbali), structurally defined ligand binding sites and

structurally defined binary domain interfaces (PIBASE, http://salilab.org/pibase)

as well as predictions of ligand binding sites, interactions between yeast

proteins, and functional consequences of human nsSNPs (LS-SNP,

http://salilab.org/LS-SNP).

DOI: 10.1093/nar/gkj059

PMCID: PMC1347422

PMID: 16381869 [Indexed for MEDLINE]

3030. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D277-80.

Flexible Structural Neighborhood--a database of protein structural similarities

and alignments.

Li Z(1), Ye Y, Godzik A.

Author information:

(1)The Burnham Institute, La Jolla, CA, USA.

Protein structures are flexible, changing their shapes not only upon substrate

binding, but also during evolution as a collective effect of mutations, deletions

and insertions. A new generation of protein structure comparison algorithms

allows for such flexibility; they go beyond identifying the largest common part

between two proteins and find hinge regions and patterns of flexibility in

protein families. Here we present a Flexible Structural Neighborhood (FSN), a

database of structural neighbors of proteins deposited in PDB as seen by a

flexible protein structure alignment program FATCAT, developed previously in our

group. The database, searchable by a protein PDB code, provides lists of proteins

with statistically significant structural similarity and on lower menu levels

provides detailed alignments, interactive superposition of structures and

positions of hinges that were identified in the comparison. While superficially

similar to other structural protein alignment resources, FSN provides a unique

resource to study not only protein structural similarity, but also how protein

structures change. FSN is available from a server

http://fatcat.burnham.org/fatcat/struct\_neighbor and by direct links from the PDB

database.

DOI: 10.1093/nar/gkj124

PMCID: PMC1347486

PMID: 16381864 [Indexed for MEDLINE]

3031. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D169-72.

MIPS: analysis and annotation of proteins from whole genomes in 2005.

Mewes HW(1), Frishman D, Mayer KF, Münsterkötter M, Noubibou O, Pagel P, Rattei

T, Oesterheld M, Ruepp A, Stümpflen V.

Author information:

(1)Institute for Bioinformatics (MIPS), GSF National Research Center for

Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany.

The Munich Information Center for Protein Sequences (MIPS at the GSF),

Neuherberg, Germany, provides resources related to genome information. Manually

curated databases for several reference organisms are maintained. Several of

these databases are described elsewhere in this and other recent NAR database

issues. In a complementary effort, a comprehensive set of >400 genomes

automatically annotated with the PEDANT system are maintained. The main goal of

our current work on creating and maintaining genome databases is to extend gene

centered information to information on interactions within a generic

comprehensive framework. We have concentrated our efforts along three lines (i)

the development of suitable comprehensive data structures and database

technology, communication and query tools to include a wide range of different

types of information enabling the representation of complex information such as

functional modules or networks Genome Research Environment System, (ii) the

development of databases covering computable information such as the basic

evolutionary relations among all genes, namely SIMAP, the sequence similarity

matrix and the CABiNet network analysis framework and (iii) the compilation and

manual annotation of information related to interactions such as protein-protein

interactions or other types of relations (e.g. MPCDB, MPPI, CYGD). All databases

described and the detailed descriptions of our projects can be accessed through

the MIPS WWW server (http://mips.gsf.de).

DOI: 10.1093/nar/gkj148

PMCID: PMC1347510

PMID: 16381839 [Indexed for MEDLINE]

3032. Promot Educ. 2006;13(4):230-5.

The Well-being Profile--an Internet tool for school health promotion.

Lintonen T(1), Konu AI.

Author information:

(1)Tampere School of Public Health, University of Tampere, Finland.

tomi.lintonen@pakk.poliisi.fi

For years, the WHO has, in accordance with the 'settings' idea, encouraged a

whole school approach when trying to promote health and well-being in schools.

This developmental study analyses the implementation and pilot phase experiences

of a holistic well-being evaluation tool for schools, the School Well-being

Profile, on the Internet. The Profile is based on the theoretically established

School Well-being Model. The School Well-being Profile consists of electronic

survey forms and an automatic facility that analyses and produces results on the

data in graphic and numeric form. After the data has been entered, the primary

user within the school can immediately view and print out the results. The

figures can be compared with the averages of all schools to pinpoint areas where

well-being is different from that in the other schools. The Profile resides on a

Finnish National Board of Education server (www2.edu.fi/hyvinvointiprofiili) and

its use is free of charge for all schools. The Profile became popular in its

first year: it was used by 33 primary schools, 28 lower secondary schools and 9

upper secondary schools with a total of 9,169 respondents. Overall, 94% of the

students and 99% of the personnel expressed that it had been at least fairly easy

to fill in the questionnaire. The paper shows that theoretical research in health

promotion can effectively be put into practice using information technology

tools. The project that produced the School Well-being Profile ended in 2004, yet

the WWW-Profile continues to gain new users. The school administrators, personnel

and students have found the Profile easy to use and worth the time and effort

invested by the school. The strategies adopted in the design and dissemination of

the Profile seem to have been successful in providing a sustainable resource to

contribute to the future promotion of well-being in schools.

PMID: 17410973 [Indexed for MEDLINE]

3033. BMC Bioinformatics. 2005 Dec 21;6:306.

WebArray: an online platform for microarray data analysis.

Xia X(1), McClelland M, Wang Y.

Author information:

(1)Genomic Core Facility, Sidney Kimmel Cancer Center, San Diego, CA 92121, USA.

xxia@skcc.org

BACKGROUND: Many cutting-edge microarray analysis tools and algorithms, including

commonly used limma and affy packages in Bioconductor, need sophisticated

knowledge of mathematics, statistics and computer skills for implementation.

Commercially available software can provide a user-friendly interface at

considerable cost. To facilitate the use of these tools for microarray data

analysis on an open platform we developed an online microarray data analysis

platform, WebArray, for bench biologists to utilize these tools to explore data

from single/dual color microarray experiments.

RESULTS: The currently implemented functions were based on limma and affy package

from Bioconductor, the spacings LOESS histogram (SPLOSH) method, PCA-assisted

normalization method and genome mapping method. WebArray incorporates these

packages and provides a user-friendly interface for accessing a wide range of key

functions of limma and others, such as spot quality weight, background

correction, graphical plotting, normalization, linear modeling, empirical bayes

statistical analysis, false discovery rate (FDR) estimation, chromosomal mapping

for genome comparison.

CONCLUSION: WebArray offers a convenient platform for bench biologists to access

several cutting-edge microarray data analysis tools. The website is freely

available at http://bioinformatics.skcc.org/webarray/. It runs on a Linux server

with Apache and MySQL.

DOI: 10.1186/1471-2105-6-306

PMCID: PMC1327694

PMID: 16371165 [Indexed for MEDLINE]

3034. BMC Bioinformatics. 2005 Dec 21;6:305.

The Gene Set Builder: collation, curation, and distribution of sets of genes.

Yusuf D(1), Lim JS, Wasserman WW.

Author information:

(1)Centre for Molecular Medicine and Therapeutics (CMMT), Child & Family Research

Institute, Vancouver, Canada. dyusuf@cmmt.ubc.ca

BACKGROUND: In bioinformatics and genomics, there are many applications designed

to investigate the common properties for a set of genes. Often, these multi-gene

analysis tools attempt to reveal sequential, functional, and expressional ties.

However, while tremendous effort has been invested in developing tools that can

analyze a set of genes, minimal effort has been invested in developing tools that

can help researchers compile, store, and annotate gene sets in the first place.

As a result, the process of making or accessing a set often involves tedious and

time consuming steps such as finding identifiers for each individual gene. These

steps are often repeated extensively to shift from one identifier type to

another; or to recreate a published set. In this paper, we present a simple

online tool which - with the help of the gene catalogs Ensembl and GeneLynx - can

help researchers build and annotate sets of genes quickly and easily.

DESCRIPTION: The Gene Set Builder is a database-driven, web-based tool designed

to help researchers compile, store, export, and share sets of genes. This

application supports the 17 eukaryotic genomes found in version 32 of the Ensembl

database, which includes species from yeast to human. User-created information

such as sets and customized annotations are stored to facilitate easy access.

Gene sets stored in the system can be "exported" in a variety of output formats -

as lists of identifiers, in tables, or as sequences. In addition, gene sets can

be "shared" with specific users to facilitate collaborations or fully released to

provide access to published results. The application also features a Perl API

(Application Programming Interface) for direct connectivity to custom analysis

tools. A downloadable Quick Reference guide and an online tutorial are available

to help new users learn its functionalities.

CONCLUSION: The Gene Set Builder is an Ensembl-facilitated online tool designed

to help researchers compile and manage sets of genes in a user-friendly

environment. The application can be accessed via http://www.cisreg.ca/gsb/.

DOI: 10.1186/1471-2105-6-305

PMCID: PMC1351202

PMID: 16371163 [Indexed for MEDLINE]

3035. Bioinformatics. 2005 Dec 1;21(23):4248-54. Epub 2005 Oct 4.

Pcons5: combining consensus, structural evaluation and fold recognition scores.

Wallner B(1), Elofsson A.

Author information:

(1)Stockholm Bioinformatics Center, Stockholm University SE-106 91 Stockholm,

Sweden. bjorn@sbc.su.se

MOTIVATION: The success of the consensus approach to the protein structure

prediction problem has led to development of several different consensus methods.

Most of them only rely on a structural comparison of a number of different

models. However, there are other types of information that might be useful such

as the score from the server and structural evaluation.

RESULTS: Pcons5 is a new and improved version of the consensus predictor Pcons.

Pcons5 integrates information from three different sources: the consensus

analysis, structural evaluation and the score from the fold recognition servers.

We show that Pcons5 is better than the previous version of Pcons and that it

performs better than using only the consensus analysis. In addition, we also

present a version of Pmodeller based on Pcons5, which performs significantly

better than Pcons5.

AVAILABILITY: Pcons5 is the first Pcons version available as a standalone program

from http://www.sbc.su.se/~bjorn/Pcons5. It should be easy to implement in local

meta-servers.

DOI: 10.1093/bioinformatics/bti702

PMID: 16204344 [Indexed for MEDLINE]

3036. BMC Bioinformatics. 2005 Dec 1;6 Suppl 4:S17.

Browsing isolated population data.

Mancosu G(1), Cosso M, Marras F, Borlino CC, Ledda G, Manias T, Adamo M, Serra D,

Melis P, Pirastu M.

Author information:

(1)SharDNA Life Sciences, Cagliari, Italy. mancosu@shardna.it

BACKGROUND: In our studies of genetically isolated populations in a remote

mountain area in the center of Sardinia (Italy), we found that 80-85% of the

inhabitants of each village belong to a single huge pedigree with families

strictly connected to each other through hundreds of loops. Moreover,

intermarriages between villages join pedigrees of different villages through

links that make family trees even more complicated. Unfortunately, none of the

commonly used pedigree drawing tools are able to draw the complete pedigree,

whereas it is commonly accepted that the visual representation of families is

very important as it helps researchers in identifying clusters of inherited

traits and genotypes. We had a representation issue that compels researchers to

work with subsets extracted from the overall genealogy, causing a serious loss of

information on familiar relationships. To visually explore such complex

pedigrees, we developed PedNavigator, a browser for genealogical databases

properly suited for genetic studies.

RESULTS: The PedNavigator is useful for genealogical research due to its capacity

to represent family relations between persons and to make a visual verification

of the links during family history reconstruction. As for genetic studies, it is

helpful to follow propagation of a specific set of genetic markers (haplotype),

or to select people for linkage analysis, showing relations between various

branch of a family tree of affected subjects.

AVAILABILITY: PedNavigator is an application integrated into a Framework designed

to handle data for human genetic studies based on the Oracle platform. To allow

the use of PedNavigator also to people not owning the same required informatics

infrastructure or systems, we developed PedNavigator Lite with mainly the same

features of the integrated one, based on MySQL database server. This version is

free for academic users, and it is available for download from our site

http://www.shardna.com.

DOI: 10.1186/1471-2105-6-S4-S17

PMCID: PMC1866391

PMID: 16351743 [Indexed for MEDLINE]

3037. BMC Bioinformatics. 2005 Dec 1;6:287.

ArrayQuest: a web resource for the analysis of DNA microarray data.

Argraves GL(1), Jani S, Barth JL, Argraves WS.

Author information:

(1)Array Genetics, Inc,, 59 Great Quarter Road, Newtown, CT 06482, USA.

argraves@musc.edu

BACKGROUND: Numerous microarray analysis programs have been created through the

efforts of Open Source software development projects. Providing browser-based

interfaces that allow these programs to be executed over the Internet enhances

the applicability and utility of these analytic software tools.

RESULTS: Here we present ArrayQuest, a web-based DNA microarray analysis process

controller. Key features of ArrayQuest are that (1) it is capable of executing

numerous analysis programs such as those written in R, BioPerl and C++; (2) new

analysis programs can be added to ArrayQuest Methods Library at the request of

users or developers; (3) input DNA microarray data can be selected from public

databases (i.e., the Medical University of South Carolina (MUSC) DNA Microarray

Database or Gene Expression Omnibus (GEO)) or it can be uploaded to the

ArrayQuest center-point web server into a password-protected area; and (4)

analysis jobs are distributed across computers configured in a backend cluster.

To demonstrate the utility of ArrayQuest we have populated the methods library

with methods for analysis of Affymetrix DNA microarray data.

CONCLUSION: ArrayQuest enables browser-based implementation of DNA microarray

data analysis programs that can be executed on a Linux-based platform.

Importantly, ArrayQuest is a platform that will facilitate the distribution and

implementation of new analysis algorithms and is therefore of use to both

developers of analysis applications as well as users. ArrayQuest is freely

available for use at http://proteogenomics.musc.edu/arrayquest.html.

DOI: 10.1186/1471-2105-6-287

PMCID: PMC1325052

PMID: 16321157 [Indexed for MEDLINE]

3038. Mol Biol Cell. 2005 Dec;16(12):5736-48. Epub 2005 Sep 29.

Sequence and comparative genomic analysis of actin-related proteins.

Muller J(1), Oma Y, Vallar L, Friederich E, Poch O, Winsor B.

Author information:

(1)Laboratoire de Biologie et Génomique Structurales, Institut de Génétique et de

Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP 163, 67404 Illkirch

Cedex, France. jean.muller@igbmc.u-strasbg.fr

Actin-related proteins (ARPs) are key players in cytoskeleton activities and

nuclear functions. Two complexes, ARP2/3 and ARP1/11, also known as dynactin, are

implicated in actin dynamics and in microtubule-based trafficking, respectively.

ARP4 to ARP9 are components of many chromatin-modulating complexes. Conventional

actins and ARPs codefine a large family of homologous proteins, the actin

superfamily, with a tertiary structure known as the actin fold. Because ARPs and

actin share high sequence conservation, clear family definition requires distinct

features to easily and systematically identify each subfamily. In this study we

performed an in depth sequence and comparative genomic analysis of ARP

subfamilies. A high-quality multiple alignment of approximately 700 complete

protein sequences homologous to actin, including 148 ARP sequences, allowed us to

extend the ARP classification to new organisms. Sequence alignments revealed

conserved residues, motifs, and inserted sequence signatures to define each ARP

subfamily. These discriminative characteristics allowed us to develop ARPAnno

(http://bips.u-strasbg.fr/ARPAnno), a new web server dedicated to the annotation

of ARP sequences. Analyses of sequence conservation among actins and ARPs

highlight part of the actin fold and suggest interactions between ARPs and

actin-binding proteins. Finally, analysis of ARP distribution across eukaryotic

phyla emphasizes the central importance of nuclear ARPs, particularly the

multifunctional ARP4.

DOI: 10.1091/mbc.E05-06-0508

PMCID: PMC1289417

PMID: 16195354 [Indexed for MEDLINE]

3039. Oral Microbiol Immunol. 2005 Dec;20(6):344-8.

A searchable database for proteomes of oral microorganisms.

Nakano Y(1), Shibata Y, Kawada M, Kojima M, Fukamachi H, Shibata Y, Okano S,

Matsushima K, Abiko Y, Yamashita Y.

Author information:

(1)Department of Preventive Dentistry, Kyushu University Faculty of Dental

Science, Fukuoka, Japan. yosh@dent.kyushu-u.ac.jp

Erratum in

Oral Microbiol Immunol. 2006 Apr;21(2):136.

An online database of proteomes for two-dimensional electrophoresis (2DE) gel

data was constructed and it is now freely accessible through a web-based

interface. Proteins from three oral bacteria, Streptococcus mutans UA159,

Actinobacillus actinomycetemcomitans HK1651, and Porphyromonas gingivalis W83,

whose genome databases are freely available, were separated by 2DE, and protein

spots were analyzed by matrix-assisted laser desorption/ionization time-of-flight

(MALDI-TOF) and identified. About 1000 spots from the gels of P. gingivalis W83

were extracted and analyzed by MALDI-TOF, and 330 proteins were identified. In

addition, 160 of 240 spots of A. actinomycetemcomitans and 158 of 356 spots of S.

mutans were identified. Information such as spot coordinates on the gels, protein

names (predicted functions), molecular weights, isoelectroric points, and links

to online databases, including Oral Pathogen Sequence Databases of the Los Alamos

National Laboratory Bioscience Division (ORALGEN) and National Center for

Biotechnology Information (NCBI) or The Institute Genomic Research (TIGR), were

stored in tables accessible through the relational database management system

MySQL on an Apache web server. To test for functionality of this database system,

responses of S. mutans to environmental changes were analyzed using the database

and 21 spots on the gel were identified as proteins whose expression had been

increased or decreased by environmental pH change without in-gel trypsin

digestion, protein extraction, or MALDI-TOF/TOF-MS (mass spectrometer) analysis.

The identified proteins are agreement with those reported in previous papers on

acid tolerance of S. mutans, demonstrating the usefulness of the system. This

database is available at

http://www.myamagu.dent.kyushu-u.ac.jp/~bioinformatics/index.html or

http://www.bipos.mascat.nihon-u.ac.jp/index.html.

DOI: 10.1111/j.1399-302X.2005.00235.x

PMID: 16238593 [Indexed for MEDLINE]

3040. Proteins. 2005 Dec 1;61(4):1075-88.

Survey of the geometric association of domain-domain interfaces.

Kim WK(1), Ison JC.

Author information:

(1)Biotechnological Centre, TU Dresden, Germany.

Considering the limited success of the most sophisticated docking methods

available and the amount of computation required for systematic docking,

cataloging all the known interfaces may be an alternative basis for the

prediction of protein tertiary and quaternary structures. We classify domain

interfaces according to the geometry of domain-domain association. By applying a

simple and efficient method called "interface tag clustering," more than 4,000

distinct types of domain interfaces are collected from Protein Quaternary

Structure Server and Protein Data Bank. Given a pair of interacting domains, we

define "face" as the set of interacting residues in each single domain and the

pair of interacting faces as an "interface." We investigate how the geometry of

interfaces relates to a network of interacting protein families, such as how many

different binding orientations are possible between two families or whether a

family uses distinct surfaces or the same surface when the family has diverse

interaction partners from various families. We show there are, on average,

1.2-1.9 different types of interfaces between interacting domains and a

significant number of family pairs associate in multiple orientations. In

general, a family tends to use distinct faces for each partner when the family

has diverse interaction partners. Each face is highly specific to its interaction

partner and the binding orientation. The relative positions of interface residues

are generally well conserved within the same type of interface even between

remote homologs. The classification result is available at

http://www.biotec.tu-dresden.de/~wkim/supplement.

Proteins 2005. 2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20693

PMID: 16247798 [Indexed for MEDLINE]

3041. Bioinformatics. 2005 Nov 15;21(22):4107-15. Epub 2005 Aug 25.

Calibrating E-values for hidden Markov models using reverse-sequence null models.

Karplus K(1), Karchin R, Shackelford G, Hughey R.

Author information:

(1)Department of Biomolecular Engineering, University of California, Santa Cruz,

95064, USA. karplus@soe.ucsc.edu

MOTIVATION: Hidden Markov models (HMMs) calculate the probability that a sequence

was generated by a given model. Log-odds scoring provides a context for

evaluating this probability, by considering it in relation to a null hypothesis.

We have found that using a reverse-sequence null model effectively removes biases

owing to sequence length and composition and reduces the number of false

positives in a database search. Any scoring system is an arbitrary measure of the

quality of database matches. Significance estimates of scores are essential,

because they eliminate model- and method-dependent scaling factors, and because

they quantify the importance of each match. Accurate computation of the

significance of reverse-sequence null model scores presents a problem, because

the scores do not fit the extreme-value (Gumbel) distribution commonly used to

estimate HMM scores' significance.

RESULTS: To get a better estimate of the significance of reverse-sequence null

model scores, we derive a theoretical distribution based on the assumption of a

Gumbel distribution for raw HMM scores and compare estimates based on this and

other distribution families. We derive estimation methods for the parameters of

the distributions based on maximum likelihood and on moment matching

(least-squares fit for Student's t-distribution). We evaluate the modeled

distributions of scores, based on how well they fit the tail of the observed

distribution for data not used in the fitting and on the effects of the improved

E-values on our HMM-based fold-recognition methods. The theoretical distribution

provides some improvement in fitting the tail and in providing fewer false

positives in the fold-recognition test. An ad hoc distribution based on assuming

a stretched exponential tail does an even better job. The use of Student's t to

model the distribution fits well in the middle of the distribution, but provides

too heavy a tail. The moment-matching methods fit the tails better than

maximum-likelihood methods.

AVAILABILITY: Information on obtaining the SAM program suite (free for academic

use), as well as a server interface, is available at

http://www.soe.ucsc.edu/research/compbio/sam.html and the open-source random

sequence generator with varying compositional biases is available at

http://www.soe.ucsc.edu/research/compbio/gen\_sequence

DOI: 10.1093/bioinformatics/bti629

PMID: 16123115 [Indexed for MEDLINE]

3042. Proteins. 2005 Nov 15;61(3):473-80.

Predicting protein secondary structure and solvent accessibility with an improved

multiple linear regression method.

Qin S(1), He Y, Pan XM.

Author information:

(1)National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese

Academy of Sciences, Beijing, People's Republic of China.

We have improved the multiple linear regression (MLR) algorithm for protein

secondary structure prediction by combining it with the evolutionary information

provided by multiple sequence alignment of PSI-BLAST. On the CB513 dataset, the

three states average overall per-residue accuracy, Q(3), reached 76.4%, while

segment overlap accuracy, SOV99, reached 73.2%, using a rigorous jackknife

procedure and the strictest reduction of eight states DSSP definition to three

states. This represents an improvement of approximately 5% on overall per-residue

accuracy compared with previous work. The relative solvent accessibility

prediction also benefited from this combination of methods. The system achieved

77.7% average jackknifed accuracy for two states prediction based on a 25%

relative solvent accessibility mode, with a Mathews' correlation coefficient of

0.548. The improved MLR secondary structure and relative solvent accessibility

prediction server is available at http://spg.biosci.tsinghua.edu.cn/.

(c) 2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20645

PMID: 16152601 [Indexed for MEDLINE]

3043. BMC Bioinformatics. 2005 Nov 3;6:264.

maxdLoad2 and maxdBrowse: standards-compliant tools for microarray experimental

annotation, data management and dissemination.

Hancock D(1), Wilson M, Velarde G, Morrison N, Hayes A, Hulme H, Wood AJ, Nashar

K, Kell DB, Brass A.

Author information:

(1)School of Computer Science, The University of Manchester, Kilburn Building,

Oxford Road, Manchester, UK. d.hancock@cs.man.ac.uk

BACKGROUND: maxdLoad2 is a relational database schema and Java application for

microarray experimental annotation and storage. It is compliant with all

standards for microarray meta-data capture; including the specification of what

data should be recorded, extensive use of standard ontologies and support for

data exchange formats. The output from maxdLoad2 is of a form acceptable for

submission to the ArrayExpress microarray repository at the European

Bioinformatics Institute. maxdBrowse is a PHP web-application that makes contents

of maxdLoad2 databases accessible via web-browser, the command-line and

web-service environments. It thus acts as both a dissemination and data-mining

tool.

RESULTS: maxdLoad2 presents an easy-to-use interface to an underlying relational

database and provides a full complement of facilities for browsing, searching and

editing. There is a tree-based visualization of data connectivity and the ability

to explore the links between any pair of data elements, irrespective of how many

intermediate links lie between them. Its principle novel features are: the

flexibility of the meta-data that can be captured, the tools provided for

importing data from spreadsheets and other tabular representations, the tools

provided for the automatic creation of structured documents, the ability to

browse and access the data via web and web-services interfaces. Within maxdLoad2

it is very straightforward to customise the meta-data that is being captured or

change the definitions of the meta-data. These meta-data definitions are stored

within the database itself allowing client software to connect properly to a

modified database without having to be specially configured. The meta-data

definitions (configuration file) can also be centralized allowing changes made in

response to revisions of standards or terminologies to be propagated to clients

without user intervention.maxdBrowse is hosted on a web-server and presents

multiple interfaces to the contents of maxd databases. maxdBrowse emulates many

of the browse and search features available in the maxdLoad2 application via a

web-browser. This allows users who are not familiar with maxdLoad2 to browse and

export microarray data from the database for their own analysis. The same browse

and search features are also available via command-line and SOAP server

interfaces. This both enables scripting of data export for use embedded in data

repositories and analysis environments, and allows access to the maxd databases

via web-service architectures.

CONCLUSION: maxdLoad2 http://www.bioinf.man.ac.uk/microarray/maxd/ and maxdBrowse

http://dbk.ch.umist.ac.uk/maxdBrowse are portable and compatible with all common

operating systems and major database servers. They provide a powerful, flexible

package for annotation of microarray experiments and a convenient dissemination

environment. They are available for download and open sourced under the Artistic

License.

DOI: 10.1186/1471-2105-6-264

PMCID: PMC1298287

PMID: 16269077 [Indexed for MEDLINE]

3044. Acta Biochim Biophys Sin (Shanghai). 2005 Nov;37(11):759-66.

Fast fourier transform-based support vector machine for prediction of G-protein

coupled receptor subfamilies.

Guo YZ(1), Li ML, Wang KL, Wen ZN, Lu MC, Liu LX, Jiang L.

Author information:

(1)College of Chemistry, Sichuan University, Chengdu 610064, China.

Erratum in

Acta Biochim Biophys Sin (Shanghai). 2006 Jul;38(7):456.

Although the sequence information on G-protein coupled receptors (GPCRs)

continues to grow, many GPCRs remain orphaned (i.e. ligand specificity unknown)

or poorly characterized with little structural information available, so an

automated and reliable method is badly needed to facilitate the identification of

novel receptors. In this study, a method of fast Fourier transform-based support

vector machine has been developed for predicting GPCR subfamilies according to

protein's hydrophobicity. In classifying Class B, C, D and F subfamilies, the

method achieved an overall Matthe's correlation coefficient and accuracy of 0.95

and 93.3%, respectively, when evaluated using the jackknife test. The method

achieved an accuracy of 100% on the Class B independent dataset. The results show

that this method can classify GPCR subfamilies as well as their functional

classification with high accuracy. A web server implementing the prediction is

available at http://chem.scu.edu.cn/blast/Pred-GPCR.

PMID: 16270155 [Indexed for MEDLINE]

3045. Bioinformatics. 2005 Nov 1;21 Suppl 3:iii20-30.

Incorporation of splice site probability models for non-canonical introns

improves gene structure prediction in plants.

Sparks ME(1), Brendel V.

Author information:

(1)Department of Genetics, Development and Cell Biology, Iowa State University,

Ames, IA 50011-3260, USA.

MOTIVATION: The vast majority of introns in protein-coding genes of higher

eukaryotes have a GT dinucleotide at their 5'-terminus and an AG dinucleotide at

their 3' end. About 1-2% of introns are non-canonical, with the most abundant

subtype of non-canonical introns being characterized by GC and AG dinucleotides

at their 5'- and 3'-termini, respectively. Most current gene prediction software,

whether based on ab initio or spliced alignment approaches, does not include

explicit models for non-canonical introns or may exclude their prediction

altogether. With present amounts of genome and transcript data, it is now

possible to apply statistical methodology to non-canonical splice site

prediction. We pursued one such approach and describe the training and

implementation of GC-donor splice site models for Arabidopsis and rice, with the

goal of exploring whether specific modeling of non-canonical introns can enhance

gene structure prediction accuracy.

RESULTS: Our results indicate that the incorporation of non-canonical splice site

models yields dramatic improvements in annotating genes containing GC-AG and

AT-AC non-canonical introns. Comparison of models shows differences between

monocot and dicot species, but also suggests GC intron-specific biases

independent of taxonomic clade. We also present evidence that GC-AG introns occur

preferentially in genes with atypically high exon counts.

AVAILABILITY: Source code for the updated versions of GeneSeqer and

SplicePredictor (distributed with the GeneSeqer code) isavailable at

http://bioinformatics.iastate.edu/bioinformatics2go/gs/download.html. Web servers

for Arabidopsis, rice and other plant species are accessible at

http://www.plantgdb.org/PlantGDB-cgi/GeneSeqer/AtGDBgs.cgi,

http://www.plantgdb.org/PlantGDB-cgi/GeneSeqer/OsGDBgs.cgi and

http://www.plantgdb.org/PlantGDB-cgi/GeneSeqer/PlantGDBgs.cgi, respectively. A

SplicePredictor web server is available at

http://bioinformatics.iastate.edu/cgi-bin/sp.cgi. Software to generate training

data and parameterizations for Bayesian splice site models is available at

http://gremlin1.gdcb.iastate.edu/~volker/SB05B/BSSM4GSQ/

DOI: 10.1093/bioinformatics/bti1205

PMID: 16306388 [Indexed for MEDLINE]

3046. Bioinformatics. 2005 Nov 1;21(21):3963-9. Epub 2005 Sep 6.

pTARGET [corrected] a new method for predicting protein subcellular localization

in eukaryotes.

Guda C(1), Subramaniam S.

Author information:

(1)Gen\*NY\*sis Center for Excellence in Cancer Genomics, State University of New

York, One Discovery Drive, Rensselaer, NY 12144-3456, USA. cguda@albany.edu

Erratum in

Bioinformatics. 2005 Dec 15;21(24):4434.

MOTIVATION: There is a scarcity of efficient computational methods for predicting

protein subcellular localization in eukaryotes. Currently available methods are

inadequate for genome-scale predictions with several limitations. Here, we

present a new prediction method, pTARGET that can predict proteins targeted to

nine different subcellular locations in the eukaryotic animal species.

RESULTS: The nine subcellular locations predicted by pTARGET include cytoplasm,

endoplasmic reticulum, extracellular/secretory, golgi, lysosomes, mitochondria,

nucleus, plasma membrane and peroxisomes. Predictions are based on the

location-specific protein functional domains and the amino acid compositional

differences across different subcellular locations. Overall, this method can

predict 68-87% of the true positives at accuracy rates of 96-99%. Comparison of

the prediction performance against PSORT showed that pTARGET prediction rates are

higher by 11-60% in 6 of the 8 locations tested. Besides, the pTARGET method is

robust enough for genome-scale prediction of protein subcellular localizations

since, it does not rely on the presence of signal or target peptides.

AVAILABILITY: A public web server based on the pTARGET method is accessible at

the URL http://bioinformatics.albany.edu/~ptarget. Datasets used for developing

pTARGET can be downloaded from this web server. Source code will be available on

request from the corresponding author.

DOI: 10.1093/bioinformatics/bti650

PMID: 16144808 [Indexed for MEDLINE]

3047. Braz J Med Biol Res. 2005 Nov;38(11):1571-4. Epub 2005 Oct 26.

The GATO gene annotation tool for research laboratories.

Fujita A(1), Massirer KB, Durham AM, Ferreira CE, Sogayar MC.

Author information:

(1)Instituto de Química, Universidade de São Paulo, São Paulo, SP, Brazil.

Large-scale genome projects have generated a rapidly increasing number of DNA

sequences. Therefore, development of computational methods to rapidly analyze

these sequences is essential for progress in genomic research. Here we present an

automatic annotation system for preliminary analysis of DNA sequences. The gene

annotation tool (GATO) is a Bioinformatics pipeline designed to facilitate

routine functional annotation and easy access to annotated genes. It was designed

in view of the frequent need of genomic researchers to access data pertaining to

a common set of genes. In the GATO system, annotation is generated by querying

some of the Web-accessible resources and the information is stored in a local

database, which keeps a record of all previous annotation results. GATO may be

accessed from everywhere through the internet or may be run locally if a large

number of sequences are going to be annotated. It is implemented in PHP and Perl

and may be run on any suitable Web server. Usually, installation and application

of annotation systems require experience and are time consuming, but GATO is

simple and practical, allowing anyone with basic skills in informatics to access

it without any special training. GATO can be downloaded at

[http://mariwork.iq.usp.br/gato/]. Minimum computer free space required is 2 MB.

DOI: /S0100-879X2005001100002

PMID: 16258624 [Indexed for MEDLINE]

3048. J Mol Model. 2005 Nov;11(6):431-8. Epub 2005 Aug 11.

Molecular modeling of phosphorylation sites in proteins using a database of local

structure segments.

Plewczynski D(1), Jaroszewski L, Godzik A, Kloczkowski A, Rychlewski L.

Author information:

(1)BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland.

darman@bioinfo.pl

A new bioinformatics tool for molecular modeling of the local structure around

phosphorylation sites in proteins has been developed. Our method is based on a

library of short sequence and structure motifs. The basic structural elements to

be predicted are local structure segments (LSSs). This enables us to avoid the

problem of non-exact local description of structures, caused by either diversity

in the structural context, or uncertainties in prediction methods. We have

developed a library of LSSs and a profile--profile-matching algorithm that

predicts local structures of proteins from their sequence information. Our

fragment library prediction method is publicly available on a server (FRAGlib),

at http://ffas.ljcrf.edu/Servers/frag.html . The algorithm has been applied

successfully to the characterization of local structure around phosphorylation

sites in proteins. Our computational predictions of sequence and structure

preferences around phosphorylated residues have been confirmed by phosphorylation

experiments for PKA and PKC kinases. The quality of predictions has been

evaluated with several independent statistical tests. We have observed a

significant improvement in the accuracy of predictions by incorporating

structural information into the description of the neighborhood of the

phosphorylated site. Our results strongly suggest that sequence information ought

to be supplemented with additional structural context information (predicted with

our segment similarity method) for more successful predictions of phosphorylation

sites in proteins.

DOI: 10.1007/s00894-005-0235-z

PMID: 16094535 [Indexed for MEDLINE]

3049. Proteins. 2005 Nov 1;61(2):318-24.

Real value prediction of solvent accessibility in proteins using multiple

sequence alignment and secondary structure.

Garg A(1), Kaur H, Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector-39A, Chandigarh, India.

The present study is an attempt to develop a neural network-based method for

predicting the real value of solvent accessibility from the sequence using

evolutionary information in the form of multiple sequence alignment. In this

method, two feed-forward networks with a single hidden layer have been trained

with standard back-propagation as a learning algorithm. The Pearson's correlation

coefficient increases from 0.53 to 0.63, and mean absolute error decreases from

18.2 to 16% when multiple-sequence alignment obtained from PSI-BLAST is used as

input instead of a single sequence. The performance of the method further

improves from a correlation coefficient of 0.63 to 0.67 when secondary structure

information predicted by PSIPRED is incorporated in the prediction. The final

network yields a mean absolute error value of 15.2% between the experimental and

predicted values, when tested on two different nonhomologous and nonredundant

datasets of varying sizes. The method consists of two steps: (1) in the first

step, a sequence-to-structure network is trained with the multiple alignment

profiles in the form of PSI-BLAST-generated position-specific scoring matrices,

and (2) in the second step, the output obtained from the first network and

PSIPRED-predicted secondary structure information is used as an input to the

second structure-to-structure network. Based on the present study, a server

SARpred (http://www.imtech.res.in/raghava/sarpred/) has been developed that

predicts the real value of solvent accessibility of residues for a given protein

sequence. We have also evaluated the performance of SARpred on 47 proteins used

in CASP6 and achieved a correlation coefficient of 0.68 and a MAE of 15.9%

between predicted and observed values.

Copyright 2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20630

PMID: 16106377 [Indexed for MEDLINE]

3050. Bioinformatics. 2005 Oct 15;21(20):3929-30. Epub 2005 Sep 1.

The SuMo server: 3D search for protein functional sites.

Jambon M(1), Andrieu O, Combet C, Deléage G, Delfaud F, Geourjon C.

Author information:

(1)Pôle BioInformatique Lyonnais, Institut de Biologie et Chimie des Protéines

UMR 5086 CNRS/UCBL, IFR 128, Lyon, France. mjambon@burnham.org

We provide the scientific community with a web server which gives access to SuMo,

a bioinformatic system for finding similarities in arbitrary 3D structures or

substructures of proteins. SuMo is based on a unique representation of

macromolecules using selected triplets of chemical groups having their own

geometry and symmetry, regardless of the restrictive notions of main chain and

lateral chains of amino acids. The heuristic for extracting similar sites was

used to drive two major large-scale approaches. First, searching for ligand

binding sites onto a query structure has been made possible by comparing the

structure against each of the ligand binding sites found in the Protein Data Bank

(PDB). Second, the reciprocal process, i.e. searching for a given 3D site of

interest among the structures of the PDB is also possible and helps detect

cross-reacting targets in drug design projects.AVAILABILITY: The web server is

freely accessible to academia through http://sumo-pbil.ibcp.fr and full support

is available from MEDIT (http://www.medit.fr).

CONTACT: mjambon@burnham.org.

DOI: 10.1093/bioinformatics/bti645

PMID: 16141250 [Indexed for MEDLINE]

3051. Bioinformatics. 2005 Oct 15;21(20):3926-8. Epub 2005 Aug 18.

Structure clustering features on the Sfold Web server.

Chan CY(1), Lawrence CE, Ding Y.

Author information:

(1)Bioinformatics Center, Wadsworth Center, New York State Department of Health,

Albany, 12208, USA. chanc@wadsworth.org

The energy landscape of RNA secondary structures is often complex, and the

Boltzmann-weighted ensemble usually contains distinct clusters. Furthermore, the

minimum free energy structure often lies outside of the cluster containing the

structure determined by comparative sequence analysis. We have developed

procedures to characterize and visualize the Boltzmann-weighted ensemble, and

have made them available on the Sfold Web server. The new features on the Web

server include clustering statistics, ensemble and cluster centroids,

multi-dimensional scaling display and energy landscape representation of the

Boltzmann-weighted ensemble.AVAILABILITY: http://sfold.wadsworth.org;

http://www.bioinfo.rpi.edu/applications/sfold

CONTACT: chanc@wadsworth.org.

DOI: 10.1093/bioinformatics/bti632

PMID: 16109749 [Indexed for MEDLINE]

3052. BMC Bioinformatics. 2005 Oct 12;6:247.

Phydbac "Gene Function Predictor": a gene annotation tool based on genomic

context analysis.

Enault F(1), Suhre K, Claverie JM.

Author information:

(1)Structural and Genomic Information, CNRS-UPR 2589, 31 chemin Joseph Aiguier,

13009 Marseille, France. enault@igs.cnrs-mrs.fr

BACKGROUND: The large amount of completely sequenced genomes allows genomic

context analysis to predict reliable functional associations between prokaryotic

proteins. Major methods rely on the fact that genes encoding physically

interacting partners or members of shared metabolic pathways tend to be proximate

on the genome, to evolve in a correlated manner and to be fused as a single

sequence in another organism.

RESULTS: The new "Gene Function Predictor", linked to the web server Phydbac

proposes putative associations between Escherichia coli K-12 proteins derived

from a combination of these methods. We show that associations made by this tool

are more accurate than linkages found in the other established databases.

Predicted assignments to GO categories, based on pre-existing functional

annotations of associated proteins are also available. This new database

currently holds 9,379 pairwise links at an expected success rate of at least 80%,

the 6,466 functional predictions to GO terms derived from these links having a

level of accuracy higher than 70%.

CONCLUSION: The "Gene Function Predictor" is an automatic tool that aims to help

biologists by providing them hypothetical functional predictions out of genomic

context characteristics. The "Gene Function predictor" is available at

http://www.igs.cnrs-mrs.fr/phydbac/indexPS.html.

DOI: 10.1186/1471-2105-6-247

PMCID: PMC1280922

PMID: 16221304 [Indexed for MEDLINE]

3053. Proteins. 2005 Oct 1;61(1):6-20.

Generation and analysis of a protein-protein interface data set with similar

chemical and spatial patterns of interactions.

Mintz S(1), Shulman-Peleg A, Wolfson HJ, Nussinov R.

Author information:

(1)Sackler Institute of Molecular Medicine, Department of Human Genetics and

Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv,

Israel.

Protein-protein interfaces are regions between 2 polypeptide chains that are not

covalently connected. Here, we have created a nonredundant interface data set

generated from all 2-chain interfaces in the Protein Data Bank. This data set is

unique, since it contains clusters of interfaces with similar shapes and spatial

organization of chemical functional groups. The data set allows statistical

investigation of similar interfaces, as well as the identification and analysis

of the chemical forces that account for the protein-protein associations. Toward

this goal, we have developed I2I-SiteEngine (Interface-to-Interface SiteEngine)

[Data set available at http://bioinfo3d.cs.tau.ac.il/Interfaces; Web server:

http://bioinfo3d.cs.tau.ac.il/I2I-SiteEngine]. The algorithm recognizes

similarities between protein-protein binding surfaces. I2I-SiteEngine is

independent of the sequence or the fold of the proteins that comprise the

interfaces. In addition to geometry, the method takes into account both the

backbone and the side-chain physicochemical properties of the interacting atom

groups. Its high efficiency makes it suitable for large-scale database searches

and classifications. Below, we briefly describe the I2I-SiteEngine method. We

focus on the classification process and the obtained nonredundant protein-protein

interface data set. In particular, we analyze the biological significance of the

clusters and present examples which illustrate that given constellations of

chemical groups in protein-protein binding sites may be preferred, and are

observed in proteins with different structures and different functions. We expect

that these would yield further information regarding the forces stabilizing

protein-protein interactions.

2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20580

PMID: 16184518 [Indexed for MEDLINE]

3054. Virus Res. 2005 Oct;113(1):64-71.

Analysis and prediction of baculovirus promoter sequences.

Xing K(1), Deng R, Wang J, Feng J, Huang M, Wang X.

Author information:

(1)State Key Laboratory for Biocontrol, School of Life Science, Sun Yat-sen

(Zhongshan) University, Guangzhou 510275, PR China.

Consensus patterns of baculovirus sequences upstream from the translational

initiation sites have been analyzed and a web tool, Local Alignment Promoter

Predictor (LAPP), for the prediction of baculovirus promoter sequences has also

been developed. Potential consensus sequences, i.e., TCATTGT, TCTTGTA, CTCGTAA,

TCCATTT and TCATT plus TCGT in approximately 30 bp spacing context, have been

found in baculovirus promoter regions, in addition to well-characterized late and

early promoter elements G/T/ATAAG and TATAA, which is accompanied about 30-bp

downstream by a transcriptional initiation sequence CAGT or CATT. Promoter

prediction is performed by a dynamic programming algorithm based on maximal

segment pair measure with scores above some cutoff against each sequence in a

refined promoter database. The algorithm was able to discriminate between

promoter and non-promoter sequences in a test set of baculovirus sequences with

prediction specificity and sensitivity superior to that using five other

eukaryotic promoter recognition programs available on the Internet. A web server

that implements the LAPP with continually updated promoter database is freely

available at http://life.zsu.edu.cn/LAPP/.

DOI: 10.1016/j.virusres.2005.04.016

PMID: 15908030 [Indexed for MEDLINE]

3055. Bioinformatics. 2005 Sep 15;21(18):3615-21. Epub 2005 Jul 14.

SPEM: improving multiple sequence alignment with sequence profiles and predicted

secondary structures.

Zhou H(1), Zhou Y.

Author information:

(1)Department of Physiology and Biophysics, Howard Hughes Medical Institute

Center for Single Molecule Biophysics, State University of New York, Buffalo, NY

14214, USA.

MOTIVATION: Multiple sequence alignment is an essential part of bioinformatics

tools for a genome-scale study of genes and their evolution relations. However,

making an accurate alignment between remote homologs is challenging. Here, we

develop a method, called SPEM, that aligns multiple sequences using pre-processed

sequence profiles and predicted secondary structures for pairwise alignment,

consistency-based scoring for refinement of the pairwise alignment and a

progressive algorithm for final multiple alignment.

RESULTS: The alignment accuracy of SPEM is compared with those of established

methods such as ClustalW, T-Coffee, MUSCLE, ProbCons and PRALINE(PSI) in easy

(homologs) and hard (remote homologs) benchmarks. Results indicate that the

average sum of pairwise alignment scores given by SPEM are 7-15% higher than

those of the methods compared in aligning remote homologs (sequence identity

<30%). Its accuracy for aligning homologs (sequence identity >30%) is

statistically indistinguishable from those of the state-of-the-art techniques

such as ProbCons or MUSCLE 6.0.

AVAILABILITY: The SPEM server and its executables are available on

http://theory.med.buffalo.edu.

DOI: 10.1093/bioinformatics/bti582

PMID: 16020471 [Indexed for MEDLINE]

3056. Bioinformatics. 2005 Sep 1;21 Suppl 2:ii47-53.

ARTS: alignment of RNA tertiary structures.

Dror O(1), Nussinov R, Wolfson H.

Author information:

(1)School of Computer Science, Raymond and Beverly Sackler Faculty of Exact

Sciences, Tel Aviv University, Tel Aviv, Israel. oranit@post.tau.ac.il

MOTIVATION: A fast growing number of non-coding RNAs have recently been

discovered to play essential roles in many cellular processes. Similar to

proteins, understanding the functions of these active RNAs requires methods for

analyzing their tertiary structures. However, in contrast to the wide range of

structure-based approaches available for proteins, there is still a lack of

methods for studying RNA structures.

RESULTS: We present a new computational method named ARTS (alignment of RNA

tertiary structures). The method compares two nucleic acid structures (RNAs or

DNAs) and detects a-priori unknown common substructures. These substructures can

be either large global folds containing hundreds and even thousands of

nucleotides or small local tertiary motifs with at least two successive base

pairs. To the best of our knowledge, this is the first method of this type. The

method is highly-efficient and was used to conduct an all-against-all comparison

of all the RNA structures currently available in the Protein Data Bank.

AVAILABILITY: The program, a web-server and supplementary information are

available on http://bioinfo3d.cs.tau.ac.il/ARTS

DOI: 10.1093/bioinformatics/bti1108

PMID: 16204124 [Indexed for MEDLINE]

3057. Bioinformatics. 2005 Sep 1;21(17):3516-23. Epub 2005 Jul 14.

Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus

structure prediction.

Reeder J(1), Giegerich R.

Author information:

(1)Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany.

jreeder@TechFak.Uni-Bielefeld.DE

MOTIVATION: The well-known Sankoff algorithm for simultaneous RNA sequence

alignment and folding is currently considered an ideal, but computationally

over-expensive method. Available tools implement this algorithm under various

pragmatic restrictions. They are still expensive to use, and it is difficult to

judge if the moderate quality of results is because of the underlying model or to

its imperfect implementation.

RESULTS: We propose to redefine the consensus structure prediction problem in a

way that does not imply a multiple sequence alignment step. For a family of RNA

sequences, our method explicitly and independently enumerates the near-optimal

abstract shape space, and predicts as the consensus an abstract shape common to

all sequences. For each sequence, it delivers the thermodynamically best

structure which has this common shape. Since the shape space is much smaller than

the structure space, and identification of common shapes can be done in linear

time (in the number of shapes considered), the method is essentially linear in

the number of sequences. Our evaluation shows that the new method compares

favorably with available alternatives.

AVAILABILITY: The new method has been implemented in the program RNAcast and is

available on the Bielefeld Bioinformatics Server.

CONTACT: jreeder@TechFak.Uni-Bielefeld.DE, robert@TechFak.Uni-Bielefeld.DE

SUPPLEMENTARY INFORMATION: Available at

http://bibiserv.techfak.uni-bielefeld.de/rnacast/supplementary.html

DOI: 10.1093/bioinformatics/bti577

PMID: 16020472 [Indexed for MEDLINE]

3058. Bioinformatics. 2005 Sep 1;21(17):3570-1. Epub 2005 Jun 30.

P-cats: prediction of catalytic residues in proteins from their tertiary

structures.

Kinoshita K(1), Ota M.

Author information:

(1)Institute of Medical Science, University of Tokyo, Shirokanedai, Minato-ku,

Tokyo 108-8639, Japan. kino@ims.u-tokyo.ac.jp

P-cats is a web server that predicts the catalytic residues in proteins from the

atomic coordinates. P-cats receives a coordinate file of the tertiary structure

and sends out analytical results via e-mail. The reply contains a summary and two

URLs to allow the user to examine the conserved residues: one for interactive

images of the prediction results and the other for a graphical view of the

multiple sequence alignment.AVAILABILITY: P-cats is freely available at

http://p-cats.hgc.jp/p-cats

CONTACT: kino@ims.u-tokyo.ac.jp

DOI: 10.1093/bioinformatics/bti561

PMID: 15994193 [Indexed for MEDLINE]

3059. Bioinformatics. 2005 Sep 1;21(17):3501-8. Epub 2005 Jun 30.

A heuristic approach for detecting RNA H-type pseudoknots.

Huang CH(1), Lu CL, Chiu HT.

Author information:

(1)Department of Biological Science and Technology, National Chiao Tung

University, Hsinchu 300, Taiwan, Republic of China.

MOTIVATION: RNA H-type pseudoknots are ubiquitous pseudoknots that are found in

almost all classes of RNA and thought to play very important roles in a variety

of biological processes. Detection of these RNA H-type pseudoknots can improve

our understanding of RNA structures and their associated functions. However, the

currently existing programs for detecting such RNA H-type pseudoknots are still

time consuming and sometimes even ineffective. Therefore, efficient and effective

tools for detecting the RNA H-type pseudoknots are needed.

RESULTS: In this paper, we have adopted a heuristic approach to develop a novel

tool, called HPknotter, for efficiently and accurately detecting H-type

pseudoknots in an RNA sequence. In addition, we have demonstrated the

applicability and effectiveness of HPknotter by testing on some sequences with

known H-type pseudoknots. Our approach can be easily extended and applied to

other classes of more general pseudoknots.

AVAILABILITY: The web server of our HPknotter is available for online analysis at

http://bioalgorithm.life.nctu.edu.tw/HPKNOTTER/ CONTACT: cllu@mail.nctu.edu.tw,

chiu@cc.nctu.edu.tw

DOI: 10.1093/bioinformatics/bti568

PMID: 15994188 [Indexed for MEDLINE]

3060. BMC Bioinformatics. 2005 Aug 25;6:207.

A protein domain interaction interface database: InterPare.

Gong S(1), Park C, Choi H, Ko J, Jang I, Lee J, Bolser DM, Oh D, Kim DS, Bhak J.

Author information:

(1)National Genome Information Center, KRIBB, Daejeon, Korea. ssgong@kaist.ac.kr

BACKGROUND: Most proteins function by interacting with other molecules. Their

interaction interfaces are highly conserved throughout evolution to avoid

undesirable interactions that lead to fatal disorders in cells. Rational drug

discovery includes computational methods to identify the interaction sites of

lead compounds to the target molecules. Identifying and classifying protein

interaction interfaces on a large scale can help researchers discover drug

targets more efficiently.

DESCRIPTION: We introduce a large-scale protein domain interaction interface

database called InterPare http://interpare.net. It contains both inter-chain

(between chains) interfaces and intra-chain (within chain) interfaces. InterPare

uses three methods to detect interfaces: 1) the geometric distance method for

checking the distance between atoms that belong to different domains, 2)

Accessible Surface Area (ASA), a method for detecting the buried region of a

protein that is detached from a solvent when forming multimers or complexes, and

3) the Voronoi diagram, a computational geometry method that uses a mathematical

definition of interface regions. InterPare includes visualization tools to

display protein interior, surface, and interaction interfaces. It also provides

statistics such as the amino acid propensities of queried protein according to

its interior, surface, and interface region. The atom coordinates that belong to

interface, surface, and interior regions can be downloaded from the website.

CONCLUSION: InterPare is an open and public database server for protein

interaction interface information. It contains the large-scale interface data for

proteins whose 3D-structures are known. As of November 2004, there were 10,583

(Geometric distance), 10,431 (ASA), and 11,010 (Voronoi diagram) entries in the

Protein Data Bank (PDB) containing interfaces, according to the above three

methods. In the case of the geometric distance method, there are 31,620

inter-chain domain-domain interaction interfaces and 12,758 intra-chain

domain-domain interfaces.

DOI: 10.1186/1471-2105-6-207

PMCID: PMC1236910

PMID: 16122378 [Indexed for MEDLINE]

3061. BMC Genet. 2005 Aug 22;6:45.

G2D: a tool for mining genes associated with disease.

Perez-Iratxeta C(1), Wjst M, Bork P, Andrade MA.

Author information:

(1)Ontario Genomics Innovation Centre, Ottawa Health Research Institute, ON K1H

8L6, Ottawa, Canada. cperez-iratxeta@ohri.ca

BACKGROUND: Human inherited diseases can be associated by genetic linkage with

one or more genomic regions. The availability of the complete sequence of the

human genome allows examining those locations for an associated gene. We

previously developed an algorithm to prioritize genes on a chromosomal region

according to their possible relation to an inherited disease using a combination

of data mining on biomedical databases and gene sequence analysis.

RESULTS: We have implemented this method as a web application in our site G2D

(Genes to Diseases). It allows users to inspect any region of the human genome to

find candidate genes related to a genetic disease of their interest. In addition,

the G2D server includes pre-computed analyses of candidate genes for 552 linked

monogenic diseases without an associated gene, and the analysis of 18 asthma

loci.

CONCLUSION: G2D can be publicly accessed at http://www.ogic.ca/projects/g2d\_2/.

DOI: 10.1186/1471-2156-6-45

PMCID: PMC1208881

PMID: 16115313 [Indexed for MEDLINE]

3062. Biochem Biophys Res Commun. 2005 Aug 19;334(1):288-92.

Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino

acid composition to predict membrane protein types.

Shen H(1), Chou KC.

Author information:

(1)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong

University, Shanghai 200030, China.

Knowledge of membrane protein type often provides crucial hints toward

determining the function of an uncharacterized membrane protein. With the

avalanche of new protein sequences emerging during the post-genomic era, it is

highly desirable to develop an automated method that can serve as a high

throughput tool in identifying the types of newly found membrane proteins

according to their primary sequences, so as to timely make the relevant

annotations on them for the reference usage in both basic research and drug

discovery. Based on the concept of pseudo-amino acid composition [K.C. Chou,

Proteins: Struct. Funct. Genet. 43 (2001) 246-255; Erratum: Proteins: Struct.

Funct. Genet. 44 (2001) 60] that has made it possible to incorporate a

considerable amount of sequence-order effects by representing a protein sample in

terms of a set of discrete numbers, a novel predictor, the so-called "optimized

evidence-theoretic K-nearest neighbor" or "OET-KNN" classifier, was proposed. It

was demonstrated via the self-consistency test, jackknife test, and independent

dataset test that the new predictor, compared with many previous ones, yielded

higher success rates in most cases. The new predictor can also be used to improve

the prediction quality for, among many other protein attributes, structural

class, subcellular localization, enzyme family class, and G-protein coupled

receptor type. The OET-KNN classifier will be available as a web-server at

http://www.pami.sjtu.edu.cn/kcchou.

DOI: 10.1016/j.bbrc.2005.06.087

PMID: 16002049 [Indexed for MEDLINE]

3063. Nucleic Acids Res. 2005 Aug 8;33(14):4496-506. Print 2005.

Refinement of docked protein-ligand and protein-DNA structures using low

frequency normal mode amplitude optimization.

Lindahl E(1), Delarue M.

Author information:

(1)Unité de Biochimie Structurale, URA 2185 du CNRS, Institut Pasteur 25 Rue du

Dr Roux, F-75015 Paris, France.

Prediction of structural changes resulting from complex formation, both in

ligands and receptors, is an important and unsolved problem in structural

biology. In this work, we use all-atom normal modes calculated with the Elastic

Network Model as a basis set to model structural flexibility during formation of

macromolecular complexes and refine the non-bonded intermolecular energy between

the two partners (protein-ligand or protein-DNA) along 5-10 of the lowest

frequency normal mode directions. The method handles motions unrelated to the

docking transparently by first applying the modes that improve non-bonded energy

most and optionally restraining amplitudes; in addition, the method can correct

small errors in the ligand position when the first six rigid-body modes are

switched on. For a test set of six protein receptors that show an open-to-close

transition when binding small ligands, our refinement scheme reduces the protein

coordinate cRMS by 0.3-3.2 A. For two test cases of DNA structures interacting

with proteins, the program correctly refines the docked B-DNA starting form into

the expected bent DNA, reducing the DNA cRMS from 8.4 to 4.8 A and from 8.7 to

5.4 A, respectively. A public web server implementation of the refinement method

is available at http://lorentz.immstr.pasteur.fr.

DOI: 10.1093/nar/gki730

PMCID: PMC1183489

PMID: 16087736 [Indexed for MEDLINE]

3064. Bioinformatics. 2005 Aug 1;21(15):3318-9. Epub 2005 May 27.

COREX/BEST server: a web browser-based program that calculates regional stability

variations within protein structures.

Vertrees J(1), Barritt P, Whitten S, Hilser VJ.

Author information:

(1)Department of Human Biological Chemistry & Genetics, Sealy Center for

Structural Biology, University of Texas Medical Branch, Galveston, TX 77555, USA.

SUMMARY: Utilizing the user-supplied coordinates of a protein structure, the

COREX/BEST Server generates a structural thermodynamic ensemble. This

conformational ensemble can then be used to calculate the regional variations in

stability of a protein structure, and the stabilities are presented in units of

energy (kcal/mol). The regional stabilities, which are calculated at the

resolution of individual residues, can be mapped onto the protein structure for

visual representation and downloaded from the site in the form of tab delimited

text. The site provides an easy to follow summary of the theoretical and

algorithmic approaches and provides links to references for more detailed

descriptions.

AVAILABILITY: The COREX/BEST Server may be accessed through a typical web browser

by visiting http://best.utmb.edu/BEST/.

DOI: 10.1093/bioinformatics/bti520

PMID: 15923205 [Indexed for MEDLINE]

3065. Bioinformatics. 2005 Aug 1;21(15):3312-3. Epub 2005 May 26.

SECISDesign: a server to design SECIS-elements within the coding sequence.

Busch A(1), Will S, Backofen R.

Author information:

(1)Friedrich-Schiller-University Jena, Institute of Computer Science,

Ernst-Abbe-Platz 2, 07743 Jena, Germany.

SUMMARY: SECISDesign is a server for the design of SECIS-elements and arbitrary

RNA-elements within the coding sequence of an mRNA. The element has to satisfy

both structure and sequence constraints. At the same time, a certain amino acid

similarity to the original protein has to be kept. The designed sequence can be

used for recombinant expression of selenoproteins in Escherichia coli.

AVAILABILITY: The server is available at

http://www.bio.inf.uni-jena.de/Software/SECISDesign/index.html.

DOI: 10.1093/bioinformatics/bti507

PMID: 15919727 [Indexed for MEDLINE]

3066. Bioinformatics. 2005 Aug 1;21(15):3322-3. Epub 2005 May 24.

Efficient recognition of folds in protein 3D structures by the improved PRIDE

algorithm.

Gáspári Z(1), Vlahovicek K, Pongor S.

Author information:

(1)Bioinformatics Group, Biological Research Center, Hungarian Academy of

Sciences, Temesvári krt., 62, Szeged, Hungary.

An improved version of the PRIDE (PRobaility of IDEntity) fold prediction

algorithm has been developed, based on more solid statistical basis, fast search

capabilities and efficient input structure processing. The new algorithm is

effective in identifying protein structures at the 'H' level of the CATH

hierarchy.AVAILABILITY: The new algorithm is integrated into the PRIDE2 web

servers at http://pride.szbk.u-szeged.hu and http://www.icgeb.org/pride.

SUPPLEMENTARY INFORMATION: Detailed documentation and performance evaluation is

available in the description section of the PRIDE2 web server.

DOI: 10.1093/bioinformatics/bti513

PMID: 15914542 [Indexed for MEDLINE]

3067. Hum Mutat. 2005 Aug;26(2):63-8.

LOVD: easy creation of a locus-specific sequence variation database using an

"LSDB-in-a-box" approach.

Fokkema IF(1), den Dunnen JT, Taschner PE.

Author information:

(1)Center of Human and Clinical Genetics, Leiden University Medical Center,

Leiden, the Netherlands.

The completion of the human genome project has initiated, as well as provided the

basis for, the collection and study of all sequence variation between

individuals. Direct access to up-to-date information on sequence variation is

currently provided most efficiently through web-based, gene-centered,

locus-specific databases (LSDBs). We have developed the Leiden Open (source)

Variation Database (LOVD) software approaching the "LSDB-in-a-Box" idea for the

easy creation and maintenance of a fully web-based gene sequence variation

database. LOVD is platform-independent and uses PHP and MySQL open source

software only. The basic gene-centered and modular design of the database follows

the recommendations of the Human Genome Variation Society (HGVS) and focuses on

the collection and display of DNA sequence variations. With minimal effort, the

LOVD platform is extendable with clinical data. The open set-up should both

facilitate and promote functional extension with scripts written by the

community. The LOVD software is freely available from the Leiden Muscular

Dystrophy pages (www.DMD.nl/LOVD/). To promote the use of LOVD, we currently

offer curators the possibility to set up an LSDB on our Leiden server.

(c) 2005 Wiley-Liss, Inc.

DOI: 10.1002/humu.20201

PMID: 15977173 [Indexed for MEDLINE]

3068. J Bioinform Comput Biol. 2005 Aug;3(4):803-19.

Protein structure and fold prediction using Tree-Augmented naïve Bayesian

classifier.

Chinnasamy A(1), Sung WK, Mittal A.

Author information:

(1)Department of Computer Science, National University of Singapore, Singapore

117543, Singapore. arun@comp.nus.edu.sg

Due to the large volume of protein sequence data, computational methods to

determine the structure class and the fold class of a protein sequence have

become essential. Several techniques based on sequence similarity, Neural

Networks, Support Vector Machines (SVMs), etc. have been applied. Since most of

these classifiers use binary classifiers for multi-classification, there may be

(N) c2 classifiers required. This paper presents a framework using the

Tree-Augmented Bayesian Networks (TAN) which performs multi-classification based

on the theory of learning Bayesian Networks and using improved feature vector

representation of (Ding et al., 2001). In order to enhance TAN's performance,

pre-processing of data is done by feature discretization and post-processing is

done by using Mean Probability Voting (MPV) scheme. The advantage of using

Bayesian approach over other learning methods is that the network structure is

intuitive. In addition, one can read off the TAN structure probabilities to

determine the significance of each feature (say, hydrophobicity) for each class,

which helps to further understand the complexity in protein structure. The

experiments on the datasets used in three prominent recent works show that our

approach is more accurate than other discriminative methods. The framework is

implemented on the BAYESPROT web server and it is available at

http://www-appn.comp.nus.edu.sg/~bioinfo/bayesprot/Default.htm. More detailed

results are also available on the above website.

PMID: 16078362 [Indexed for MEDLINE]

3069. Protein Eng Des Sel. 2005 Aug;18(8):365-8. Epub 2005 Jun 24.

CTKPred: an SVM-based method for the prediction and classification of the

cytokine superfamily.

Huang N(1), Chen H, Sun Z.

Author information:

(1)Institute of Bioinformatics and System Biology, MOE Key Laboratory of

Bioinfomatics, State Key Laboratory of Biomembrane and Membrane Biotechnology,

Department of Biological Science and Biotechnology, Tsinghua University, Beijing

100084, China.

Cell proliferation, differentiation and death are controlled by a multitude of

cell-cell signals and loss of this control has devastating consequences.

Prominent among these regulatory signals is the cytokine superfamily, which has

crucial functions in the development, differentiation and regulation of immune

cells. In this study, a support vector machine (SVM)-based method was developed

for predicting families and subfamilies of cytokines using dipeptide composition.

The taxonomy of the cytokine superfamily with which our method complies was

described in the Cytokine Family cDNA Database (dbCFC) and the dataset used in

this study for training and testing was obtained from the dbCFC and Structural

Classification of Proteins (SCOP). The method classified cytokines and

non-cytokines with an accuracy of 92.5% by 7-fold cross-validation. The method is

further able to predict seven major classes of cytokine with an overall accuracy

of 94.7%. A server for recognition and classification of cytokines based on

multi-class SVMs has been set up at

http://bioinfo.tsinghua.edu.cn/~huangni/CTKPred/.

DOI: 10.1093/protein/gzi041

PMID: 15980017 [Indexed for MEDLINE]

3070. Protein Sci. 2005 Aug;14(8):2132-40. Epub 2005 Jun 29.

Integrated modeling of the major events in the MHC class I antigen processing

pathway.

Dönnes P(1), Kohlbacher O.

Author information:

(1)Department for Simulation of Biological Systems, WSI/ZBIT, Eberhard Karls

University Tübingen, D-72076 Tübingen, Germany.

doennes@informatik.uni-tuebingen.de

Rational design of epitope-driven vaccines is a key goal of immunoinformatics.

Typically, candidate selection relies on the prediction of MHC-peptide binding

only, as this is known to be the most selective step in the MHC class I antigen

processing pathway. However, proteasomal cleavage and transport by the

transporter associated with antigen processing (TAP) are essential steps in

antigen processing as well. While prediction methods exist for the individual

steps, no method has yet offered an integrated prediction of all three major

processing events. Here we present WAPP, a method combining prediction of

proteasomal cleavage, TAP transport, and MHC binding into a single prediction

system. The proteasomal cleavage site prediction employs a new matrix-based

method that is based on experimentally verified proteasomal cleavage sites.

Support vector regression is used for predicting peptides transported by TAP. MHC

binding is the last step in the antigen processing pathway and was predicted

using a support vector machine method, SVMHC. The individual methods are combined

in a filtering approach mimicking the natural processing pathway. WAPP thus

predicts peptides that are cleaved by the proteasome at the C terminus,

transported by TAP, and show significant affinity to MHC class I molecules. This

results in a decrease in false positive rates compared to MHC binding prediction

alone. Compared to prediction of MHC binding only, we report an increased overall

accuracy and a lower rate of false positive predictions for the HLA-A\*0201,

HLA-B\*2705, HLA-A\*01, and HLA-A\*03 alleles using WAPP. The method is available

online through our prediction server at

http://www-bs.informatik.uni-tuebingen.de/WAPP

DOI: 10.1110/ps.051352405

PMCID: PMC2279325

PMID: 15987883 [Indexed for MEDLINE]

3071. Proteins. 2005 Aug 1;60(2):224-31.

Geometry-based flexible and symmetric protein docking.

Schneidman-Duhovny D(1), Inbar Y, Nussinov R, Wolfson HJ.

Author information:

(1)School of Computer Science, Beverly and Raymond Sackler Faculty of Exact

Sciences, Tel Aviv University, Tel Aviv, Israel.

We present a set of geometric docking algorithms for rigid, flexible, and cyclic

symmetry docking. The algorithms are highly efficient and have demonstrated very

good performance in CAPRI Rounds 3-5. The flexible docking algorithm, FlexDock,

is unique in its ability to handle any number of hinges in the flexible molecule,

without degradation in run-time performance, as compared to rigid docking. The

algorithm for reconstruction of cyclically symmetric complexes successfully

assembles multimolecular complexes satisfying C(n) symmetry for any n in a matter

of minutes on a desktop PC. Most of the algorithms presented here are available

at the Tel Aviv University Structural Bioinformatics Web server

(http://bioinfo3d.cs.tau.ac.il/).

DOI: 10.1002/prot.20562

PMID: 15981269 [Indexed for MEDLINE]

3072. Proteins. 2005 Aug 1;60(2):296-301.

Development and testing of an automated approach to protein docking.

Tovchigrechko A(1), Vakser IA.

Author information:

(1)Bioinformatics Laboratory, Department of Applied Mathematics and Statistics,

State University of New York at Stony Brook, Stony Brook, New York, USA.

A new version of GRAMM was applied to Targets 14, 18, and 19 in CAPRI Round 5.

The predictions were generated without manual intervention. Ten top-ranked

matches for each target were submitted. The docking was performed by a rigid-body

procedure with a smoothed potential function to accommodate conformational

changes. The first stage was a global search on a fine grid with a projection of

a smoothed Lennard-Jones potential. The top predictions from the first stage were

subjected to the conjugate gradient minimization with the same smoothed

potential. The resulting local minima were reranked according to the weighted sum

of Lennard-Jones potential, pairwise residue-residue statistical preferences,

cluster occupancy, and the degree of the evolutionary conservation of the

predicted interface. For Targets 14 and 18, the conformation of the complex was

predicted with root-mean-square deviation (RMSD) of the ligand interface atoms

0.68 A and 1.88 A correspondingly. For Target 19, the interface areas on both

proteins were correctly predicted. The performance of the procedure was also

analyzed on the benchmark of bound-unbound protein complexes. The results show

that, on average, conformations of only 3 side-chains need to be optimized during

docking of unbound structures before the backbone changes become a limiting

factor. The GRAMM-X docking server is available for public use at

http://www.bioinformatics.ku.edu.

DOI: 10.1002/prot.20573

PMID: 15981259 [Indexed for MEDLINE]

3073. RNA. 2005 Aug;11(8):1157-66.

RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble.

Ding Y(1), Chan CY, Lawrence CE.

Author information:

(1)Bioinformatics Center, Wadsworth Center, New York State Department of Health,

Albany, NY 12208, USA. yding@wadsworth.org

Prediction of RNA secondary structure by free energy minimization has been the

standard for over two decades. Here we describe a novel method that forsakes this

paradigm for predictions based on Boltzmann-weighted structure ensemble. We

introduce the notion of a centroid structure as a representative for a set of

structures and describe a procedure for its identification. In comparison with

the minimum free energy (MFE) structure using diverse types of structural RNAs,

the centroid of the ensemble makes 30.0% fewer prediction errors as measured by

the positive predictive value (PPV) with marginally improved sensitivity. The

Boltzmann ensemble can be separated into a small number (3.2 on average) of

clusters. Among the centroids of these clusters, the "best cluster centroid" as

determined by comparison to the known structure simultaneously improves PPV by

46.5% and sensitivity by 21.7%. For 58% of the studied sequences for which the

MFE structure is outside the cluster containing the best centroid, the

improvements by the best centroid are 62.5% for PPV and 31.4% for sensitivity.

These results suggest that the energy well containing the MFE structure under the

current incomplete energy model is often different from the one for the

unavailable complete model that presumably contains the unique native structure.

Centroids are available on the Sfold server at http://sfold.wadsworth.org.

DOI: 10.1261/rna.2500605

PMCID: PMC1370799

PMID: 16043502 [Indexed for MEDLINE]

3074. Sci China C Life Sci. 2005 Aug;48(4):394-405.

A seqlet-based maximum entropy Markov approach for protein secondary structure

prediction.

Dong Q(1), Wang X, Lin L, Guan Y.

Author information:

(1)School of Computer Science and Technology, Harbin Institute of Technology,

China. qwdong@insun.hit.edu.cn

A novel method for predicting the secondary structures of proteins from amino

acid sequence has been presented. The protein secondary structure seqlets that

are analogous to the words in natural language have been extracted. These seqlets

will capture the relationship between amino acid sequence and the secondary

structures of proteins and further form the protein secondary structure

dictionary. To be elaborate, the dictionary is organism-specific. Protein

secondary structure prediction is formulated as an integrated word segmentation

and part of speech tagging problem. The word-lattice is used to represent the

results of the word segmentation and the maximum entropy model is used to

calculate the probability of a seqlet tagged as a certain secondary structure

type. The method is markovian in the seqlets, permitting efficient exact

calculation of the posterior probability distribution over all possible word

segmentations and their tags by viterbi algorithm. The optimal segmentations and

their tags are computed as the results of protein secondary structure prediction.

The method is applied to predict the secondary structures of proteins of four

organisms respectively and compared with the PHD method. The results show that

the performance of this method is higher than that of PHD by about 3.9% Q3

accuracy and 4.6% SOV accuracy. Combining with the local similarity protein

sequences that are obtained by BLAST can give better prediction. The method is

also tested on the 50 CASP5 target proteins with Q3 accuracy 78.9% and SOV

accuracy 77.1%. A web server for protein secondary structure prediction has been

constructed which is available at

http://www.insun.hit.edu.cn:81/demos/biology/index.html.

PMID: 16248433 [Indexed for MEDLINE]

3075. Clin Chim Acta. 2005 Jul 24;357(2):173-9.

Mathematical modeling of cancer: the future of prognosis and treatment.

Quaranta V(1), Weaver AM, Cummings PT, Anderson AR.

Author information:

(1)Department of Cancer Biology, Vanderbilt University School of Medicine,

Nashville, TN, USA. vito.quaranta@vanderbilt.edu

BACKGROUND: Cancer research has undergone radical changes in the past few years.

Producing information both at the basic and clinical levels is no longer the

issue. Rather, how to handle this information has become the major obstacle to

progress. Intuitive approaches are no longer feasible. The next big step will be

to implement mathematical modeling approaches to interrogate the enormous amount

of data being produced and extract useful answers (a "top-down" approach to

biology and medicine).

METHODS: Quantitative simulation of clinically relevant cancer situations-based

on experimentally validated mathematical modeling-provides an opportunity for the

researcher, and eventually the clinician, to address data and information in the

context of well-formulated questions and "what if" scenarios.

RESULTS AND CONCLUSIONS: At the Vanderbilt Integrative Cancer Biology Center

(VICBC), we are integrating cancer researchers, oncologists, chemical and

biological engineers, computational biologists, computer modelers, theoretical

and applied mathematicians, and imaging scientists, in order to implement a

vision for a combined web site and computational server that will be a home for

our mathematical modeling of cancer invasion. The web site

(www.vanderbilt.edu/VICBC/) will serve as a portal to our code, which simulates

tumor growth by calculating the dynamics of individual cancer cells (an

experimental "bottom-up" approach to complement the top-down model). Eventually,

cancer researchers outside of Vanderbilt will be able to initiate a simulation

based on providing individual cell data through a web page. We envision placing

the web site and computer cluster directly in the hands of biological researchers

involved in data mining and mathematical modeling. Furthermore, the web site will

also contain teaching props for a new generation of biomedical researchers fluent

in both mathematics and biology. This is unconventional bioinformatics: We will

be incorporating biological data and functional information into a unified

community-based mathematical framework. The result will be a tool for cancer

modeling that will ultimately have basic research, therapeutic and educational

value.

DOI: 10.1016/j.cccn.2005.03.023

PMID: 15907826 [Indexed for MEDLINE]

3076. BMC Bioinformatics. 2005 Jul 2;6:167.

Prediction of twin-arginine signal peptides.

Bendtsen JD(1), Nielsen H, Widdick D, Palmer T, Brunak S.

Author information:

(1)Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University

of Denmark, Building 208, DK-2800, Lyngby, Denmark. jannick@cbs.dtu.dk

BACKGROUND: Proteins carrying twin-arginine (Tat) signal peptides are exported

into the periplasmic compartment or extracellular environment independently of

the classical Sec-dependent translocation pathway. To complement other methods

for classical signal peptide prediction we here present a publicly available

method, TatP, for prediction of bacterial Tat signal peptides.

RESULTS: We have retrieved sequence data for Tat substrates in order to train a

computational method for discrimination of Sec and Tat signal peptides. The TatP

method is able to positively classify 91% of 35 known Tat signal peptides and 84%

of the annotated cleavage sites of these Tat signal peptides were correctly

predicted. This method generates far less false positive predictions on various

datasets than using simple pattern matching. Moreover, on the same datasets TatP

generates less false positive predictions than a complementary rule based

prediction method.

CONCLUSION: The method developed here is able to discriminate Tat signal peptides

from cytoplasmic proteins carrying a similar motif, as well as from Sec signal

peptides, with high accuracy. The method allows filtering of input sequences

based on Perl syntax regular expressions, whereas hydrophobicity discrimination

of Tat- and Sec-signal peptides is carried out by an artificial neural network. A

potential cleavage site of the predicted Tat signal peptide is also reported. The

TatP prediction server is available as a public web server at

http://www.cbs.dtu.dk/services/TatP/.

DOI: 10.1186/1471-2105-6-167

PMCID: PMC1182353

PMID: 15992409 [Indexed for MEDLINE]

3077. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W331-6.

PRISM: protein interactions by structural matching.

Ogmen U(1), Keskin O, Aytuna AS, Nussinov R, Gursoy A.

Author information:

(1)Koc University, Center for Computational Biology and Bioinformatics and

College of Engineering Rumelifeneri Yolu, Sariyer, Istanbul 34450, Turkey.

Prism (http://gordion.hpc.eng.ku.edu.tr/prism) is a website for protein interface

analysis and prediction of putative protein-protein interactions. It is composed

of a database holding protein interface structures derived from the Protein Data

Bank (PDB). The server also includes summary information about related proteins

and an interactive protein interface viewer. A list of putative protein-protein

interactions obtained by running our prediction algorithm can also be accessed.

These results are applied to a set of protein structures obtained from the PDB at

the time of algorithm execution (January 2004). Users can browse through the

non-redundant dataset of representative interfaces on which the prediction

algorithm depends, retrieve the list of similar structures to these interfaces or

see the results of interaction predictions for a particular protein. Another

service provided is interactive prediction. This is done by running the algorithm

for user input structures.

PMCID: PMC1160261

PMID: 15991339 [Indexed for MEDLINE]

3078. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W271-3.

SVC: structured visualization of evolutionary sequence conservation.

Roepcke S(1), Fiziev P, Seeburg PH, Vingron M.

Author information:

(1)Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin,

Germany. roepcke@molgen.mpg.de

We have developed a web application for the detailed analysis and visualization

of evolutionary sequence conservation in complex vertebrate genes. Given a pair

of orthologous genes, the protein-coding sequences are aligned. When these

sequences are mapped back onto their encoding exons in the genomes, a scaffold of

the conserved gene structure naturally emerges. Sequence similarity between exons

and introns is analysed and embedded into the gene structure scaffold. The

visualization on the SVC server provides detailed information about

evolutionarily conserved features of these genes. It further allows concise

representation of complex splice patterns in the context of evolutionary

conservation. A particular application of our tool arises from the fact that

around mRNA editing sites both exonic and intronic sequences are highly

conserved. This aids in delineation of these sites. SVC is available at

http://svc.molgen.mpg.de.

PMCID: PMC1160265

PMID: 15991338 [Indexed for MEDLINE]

3079. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W233-8.

MEDock: a web server for efficient prediction of ligand binding sites based on a

novel optimization algorithm.

Chang DT(1), Oyang YJ, Lin JH.

Author information:

(1)Department of Computer Science and Information Engineering, National Taiwan

University Taipei 106, Taiwan, ROC.

The prediction of ligand binding sites is an essential part of the drug discovery

process. Knowing the location of binding sites greatly facilitates the search for

hits, the lead optimization process, the design of site-directed mutagenesis

experiments and the hunt for structural features that influence the selectivity

of binding in order to minimize the drug's adverse effects. However, docking is

still the rate-limiting step for such predictions; consequently, much more

efficient algorithms are required. In this article, the design of the MEDock web

server is described. The goal of this sever is to provide an efficient utility

for predicting ligand binding sites. The MEDock web server incorporates a global

search strategy that exploits the maximum entropy property of the Gaussian

probability distribution in the context of information theory. As a result of the

global search strategy, the optimization algorithm incorporated in MEDock is

significantly superior when dealing with very rugged energy landscapes, which

usually have insurmountable barriers. This article describes four different

benchmark cases that span a diverse set of different types of ligand binding

interactions. These benchmarks were compared with the use of the Lamarckian

genetic algorithm (LGA), which is the major workhorse of the well-known AutoDock

program. These results demonstrate that MEDock consistently converged to the

correct binding modes with significantly smaller numbers of energy evaluations

than the LGA required. When judged by a threshold of the number of energy

evaluations consumed in the docking simulation, MEDock also greatly elevates the

rate of accurate predictions for all benchmark cases. MEDock is available at

http://medock.csie.ntu.edu.tw/ and http://bioinfo.mc.ntu.edu.tw/medock/.

PMCID: PMC1160262

PMID: 15991337 [Indexed for MEDLINE]

3080. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W202-7.

Pcleavage: an SVM based method for prediction of constitutive proteasome and

immunoproteasome cleavage sites in antigenic sequences.

Bhasin M(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector 39-A, Chandigarh, 160036, India.

This manuscript describes a support vector machine based method for the

prediction of constitutive as well as immunoproteasome cleavage sites in

antigenic sequences. This method achieved Matthew's correlation coefficents of

0.54 and 0.43 on in vitro and major histocompatibility complex ligand data,

respectively. This shows that the performance of our method is comparable to that

of the NetChop method, which is currently considered to be the best method for

proteasome cleavage site prediction. Based on the method, a web server,

Pcleavage, has also been developed. This server accepts protein sequences in any

standard format and present results in a user-friendly format. The server is

available for free use by all academic users at the URL

http://www.imtech.res.in/raghava/pcleavage/ or

http://bioinformatics.uams.edu/mirror/pcleavage/.

PMCID: PMC1160263

PMID: 15988831 [Indexed for MEDLINE]

3081. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W154-9.

BhairPred: prediction of beta-hairpins in a protein from multiple alignment

information using ANN and SVM techniques.

Kumar M(1), Bhasin M, Natt NK, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology Sector 39-A,

Chandigarh, India.

This paper describes a method for predicting a supersecondary structural motif,

beta-hairpins, in a protein sequence. The method was trained and tested on a set

of 5102 hairpins and 5131 non-hairpins, obtained from a non-redundant dataset of

2880 proteins using the DSSP and PROMOTIF programs. Two machine-learning

techniques, an artificial neural network (ANN) and a support vector machine

(SVM), were used to predict beta-hairpins. An accuracy of 65.5% was achieved

using ANN when an amino acid sequence was used as the input. The accuracy

improved from 65.5 to 69.1% when evolutionary information (PSI-BLAST profile),

observed secondary structure and surface accessibility were used as the inputs.

The accuracy of the method further improved from 69.1 to 79.2% when the SVM was

used for classification instead of the ANN. The performances of the methods

developed were assessed in a test case, where predicted secondary structure and

surface accessibility were used instead of the observed structure. The highest

accuracy achieved by the SVM based method in the test case was 77.9%. A maximum

accuracy of 71.1% with Matthew's correlation coefficient of 0.41 in the test case

was obtained on a dataset previously used by X. Cruz, E. G. Hutchinson, A.

Shephard and J. M. Thornton (2002) Proc. Natl Acad. Sci. USA, 99, 11157-11162.

The performance of the method was also evaluated on proteins used in the '6th

community-wide experiment on the critical assessment of techniques for protein

structure prediction (CASP6)'. Based on the algorithm described, a web server,

BhairPred (http://www.imtech.res.in/raghava/bhairpred/), has been developed,

which can be used to predict beta-hairpins in a protein using the SVM approach.

PMCID: PMC1160264

PMID: 15988830 [Indexed for MEDLINE]

3082. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W94-8.

PISCES: recent improvements to a PDB sequence culling server.

Wang G(1), Dunbrack RL Jr.

Author information:

(1)Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue,

Philadelphia, PA 19111, USA.

PISCES is a database server for producing lists of sequences from the Protein

Data Bank (PDB) using a number of entry- and chain-specific criteria and mutual

sequence identity. Our goal in culling the PDB is to provide the longest list

possible of the highest resolution structures that fulfill the sequence identity

and structural quality cut-offs. The new PISCES server uses a combination of

PSI-BLAST and structure-based alignments to determine sequence identities.

Structure alignment produces more complete alignments and therefore more accurate

sequence identities than PSI-BLAST. PISCES now allows a user to cull the PDB

by-entry in addition to the standard culling by individual chains. In this

scenario, a list will contain only entries that do not have a chain that has a

sequence identity to any chain in any other entry in the list over the sequence

identity cut-off. PISCES also provides fully annotated sequences including gene

name and species. The server allows a user to cull an input list of entries or

chains, so that other criteria, such as function, can be used. Results from a

search on the re-engineered RCSB's site for the PDB can be entered into the

PISCES server by a single click, combining the powerful searching abilities of

the PDB with PISCES's utilities for sequence culling. The server's data are

updated weekly. The server is available at http://dunbrack.fccc.edu/pisces.

DOI: 10.1093/nar/gki402

PMCID: PMC1160163

PMID: 15980589 [Indexed for MEDLINE]

3083. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W89-93.

ProFunc: a server for predicting protein function from 3D structure.

Laskowski RA(1), Watson JD, Thornton JM.

Author information:

(1)European Bioinformatics Institute Wellcome Trust Genome Campus, Hinxton,

Cambridge CB10 1SD, UK. roman@ebi.ac.uk

ProFunc (http://www.ebi.ac.uk/thornton-srv/databases/ProFunc) is a web server for

predicting the likely function of proteins whose 3D structure is known but whose

function is not. Users submit the coordinates of their structure to the server in

PDB format. ProFunc makes use of both existing and novel methods to analyse the

protein's sequence and structure identifying functional motifs or close

relationships to functionally characterized proteins. A summary of the analyses

provides an at-a-glance view of what each of the different methods has found.

More detailed results are available on separate pages. Often where one method has

failed to find anything useful another may be more forthcoming. The server is

likely to be of most use in structural genomics where a large proportion of the

proteins whose structures are solved are of hypothetical proteins of unknown

function. However, it may also find use in a comparative analysis of members of

large protein families. It provides a convenient compendium of sequence and

structural information that often hold vital functional clues to be followed up

experimentally.

DOI: 10.1093/nar/gki414

PMCID: PMC1160175

PMID: 15980588 [Indexed for MEDLINE]

3084. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W85-8.

Fragment Finder: a web-based software to identify similar three-dimensional

structural motif.

Ananthalakshmi P(1), Kumar ChK, Jeyasimhan M, Sumathi K, Sekar K.

Author information:

(1)Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India.

FF (Fragment Finder) is a web-based interactive search engine developed to

retrieve the user-desired similar 3D structural fragments from the selected

subset of 25 or 90% non-homologous protein chains. The search is based on the

comparison of the main chain backbone conformational angles (phi and ).

Additionally, the queried motifs can be superimposed to find out how similar the

structural fragments are, so that the information can be effectively used in

molecular modeling. The engine has facilities to view the resultant superposed or

individual 3D structure(s) on the client machine. The proposed web server is made

freely accessible at the following URL: http://cluster.physics.iisc.ernet.in/ff/

or http://144.16.71.148/ff/.

DOI: 10.1093/nar/gki353

PMCID: PMC1160114

PMID: 15980587 [Indexed for MEDLINE]

3085. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W783-6.

GoPubMed: exploring PubMed with the Gene Ontology.

Doms A(1), Schroeder M.

Author information:

(1)Biotechnological Center, Department of Computer Science TU, Dresden, Germany.

The biomedical literature grows at a tremendous rate and PubMed comprises already

over 15 000 000 abstracts. Finding relevant literature is an important and

difficult problem. We introduce GoPubMed, a web server which allows users to

explore PubMed search results with the Gene Ontology (GO), a hierarchically

structured vocabulary for molecular biology. GoPubMed provides the following

benefits: first, it gives an overview of the literature abstracts by categorizing

abstracts according to the GO and thus allowing users to quickly navigate through

the abstracts by category. Second, it automatically shows general ontology terms

related to the original query, which often do not even appear directly in the

abstract. Third, it enables users to verify its classification because GO terms

are highlighted in the abstracts and as each term is labelled with an accuracy

percentage. Fourth, exploring PubMed abstracts with GoPubMed is useful as it

shows definitions of GO terms without the need for further look up. GoPubMed is

online at www.gopubmed.org. Querying is currently limited to 100 papers per

query.

DOI: 10.1093/nar/gki470

PMCID: PMC1160231

PMID: 15980585 [Indexed for MEDLINE]

3086. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W77-80.

PROTINFO: new algorithms for enhanced protein structure predictions.

Hung LH(1), Ngan SC, Liu T, Samudrala R.

Author information:

(1)Computational Genomics Group, Department of Microbiology, University of

Washington School of Medicine, Seattle, WA 98195, USA.

We describe new algorithms and modules for protein structure prediction available

as part of the PROTINFO web server. The modules, comparative and de novo

modelling, have significantly improved back-end algorithms that were rigorously

evaluated at the sixth meeting on the Critical Assessment of Protein Structure

Prediction methods. We were one of four server groups invited to make an oral

presentation (only the best performing groups are asked to do so). These two

modules allow a user to submit a protein sequence and return atomic coordinates

representing the tertiary structure of that protein. The PROTINFO server is

available at http://protinfo.compbio.washington.edu.

DOI: 10.1093/nar/gki403

PMCID: PMC1160164

PMID: 15980581 [Indexed for MEDLINE]

3087. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W766-9.

MRS: a fast and compact retrieval system for biological data.

Hekkelman ML(1), Vriend G.

Author information:

(1)Centre for Molecular and Biomolecular Informatics, University of Nijmegen,

Toernooiveld 1, 6525 ED Nijmegen, The Netherlands. m.hekkelman@cmbi.ru.nl

The biological data explosion of the 'omics' era requires fast access to many

data types in rapidly growing data banks. The MRS server allows for very rapid

queries in a large number of flat-file data banks, such as EMBL, UniProt, OMIM,

dbEST, PDB, KEGG, etc. This server combines a fast and reliable backend with a

very user-friendly implementation of all the commonly used information retrieval

facilities. The MRS server is freely accessible at http://mrs.cmbi.ru.nl/.

Moreover, the MRS software is freely available at http://mrs.cmbi.ru.nl/download/

for those interested in making their own data banks available via a web-based

server.

DOI: 10.1093/nar/gki422

PMCID: PMC1160183

PMID: 15980580 [Indexed for MEDLINE]

3088. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W758-61.

GeneSeeker: extraction and integration of human disease-related information from

web-based genetic databases.

van Driel MA(1), Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G.

Author information:

(1)Centre for Molecular and Biomolecular Informatics, Radboud University

Nijmegen, PO Box 9010, 6500GL Nijmegen, The Netherlands.

The identification of genes underlying human genetic disorders requires the

combination of data related to cytogenetic localization, phenotypes and

expression patterns, to generate a list of candidate genes. In the field of human

genetics, it is normal to perform this combination analysis by hand. We report on

GeneSeeker (http://www.cmbi.ru.nl/GeneSeeker/), a web server that gathers and

combines data from a series of databases. All database searches are performed via

the web interfaces provided with the original databases, guaranteeing that the

most recent data are queried, and obviating data warehousing. GeneSeeker makes

the same selection of candidate genes as the human geneticists would have

performed, and thus reducing the time-consuming process to a few minutes.

GeneSeeker is particularly well suited for syndromes in which the disease gene

displays altered expression patterns in the affected tissue(s).

DOI: 10.1093/nar/gki435

PMCID: PMC1160196

PMID: 15980578 [Indexed for MEDLINE]

3089. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W72-6.

SCRATCH: a protein structure and structural feature prediction server.

Cheng J(1), Randall AZ, Sweredoski MJ, Baldi P.

Author information:

(1)Institute for Genomics and Bioinformatics, University of California Irvine,

CA, USA.

SCRATCH is a server for predicting protein tertiary structure and structural

features. The SCRATCH software suite includes predictors for secondary structure,

relative solvent accessibility, disordered regions, domains, disulfide bridges,

single mutation stability, residue contacts versus average, individual residue

contacts and tertiary structure. The user simply provides an amino acid sequence

and selects the desired predictions, then submits to the server. Results are

emailed to the user. The server is available at

http://www.igb.uci.edu/servers/psss.html.

DOI: 10.1093/nar/gki396

PMCID: PMC1160157

PMID: 15980571 [Indexed for MEDLINE]

3090. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W717-23.

GFINDer: genetic disease and phenotype location statistical analysis and mining

of dynamically annotated gene lists.

Masseroli M(1), Galati O, Pinciroli F.

Author information:

(1)BioMedical Informatics Laboratory, Bioengineering Department Politecnico di

Milano, I-20133 Milano, Italy. masseroli@biomed.polimi.it

Phenotype analysis is commonly recognized to be of great importance for gaining

insight into genetic interaction underlying inherited diseases. However, few

computational contributions have been proposed for this purpose, mainly owing to

lack of controlled clinical information easily accessible and structured for

computational genome-wise analyses. We developed and made available through

GFINDer web server an original approach for the analysis of genetic disorder

related genes by exploiting the information on genetic diseases and their

clinical phenotypes present in textual form within the Online Mendelian

Inheritance in Man (OMIM) database. Because several synonyms for the same name

and different names for overlapping concepts are often used in OMIM, we first

normalized phenotype location descriptions reducing them to a list of unique

controlled terms representing phenotype location categories. Then, we

hierarchically structured them and the correspondent genetic diseases according

to their topology and granularity of description, respectively. Thus, in GFINDer

we could implement specific Genetic Disorders modules for the analysis of these

structured data. Such modules allow to automatically annotate user-classified

gene lists with updated disease and clinical information, classify them according

to the genetic syndrome and the phenotypic location categories, and statistically

identify the most relevant categories in each gene class. GFINDer is available

for non-profit use at http://www.bioinformatics.polimi.it/GFINDer/.

DOI: 10.1093/nar/gki454

PMCID: PMC1160215

PMID: 15980570 [Indexed for MEDLINE]

3091. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W710-6.

BRIGEP--the BRIDGE-based genome-transcriptome-proteome browser.

Goesmann A(1), Linke B, Bartels D, Dondrup M, Krause L, Neuweger H, Oehm S,

Paczian T, Wilke A, Meyer F.

Author information:

(1)Bielefeld University, Center for Biotechnology (CeBiTec), D-33594 Bielefeld,

Germany. Alexander.Goesmann@CeBiTec.Uni-Bielefeld.DE

The growing amount of information resulting from the increasing number of

publicly available genomes and experimental results thereof necessitates the

development of comprehensive systems for data processing and analysis. In this

paper, we describe the current state and latest developments of our BRIGEP

bioinformatics software system consisting of three web-based applications: GenDB,

EMMA and ProDB. These applications facilitate the processing and analysis of

bacterial genome, transcriptome and proteome data and are actively used by

numerous international groups. We are currently in the process of extensively

interconnecting these applications. BRIGEP was developed in the Bioinformatics

Resource Facility of the Center for Biotechnology at Bielefeld University and is

freely available. A demo project with sample data and access to all three tools

is available at https://www.cebitec.uni-bielefeld.de/groups/brf/software/brigep/.

Code bundles for these and other tools developed in our group are accessible on

our FTP server at ftp.cebitec.uni-bielefeld.de/pub/software/.

DOI: 10.1093/nar/gki400

PMCID: PMC1160161

PMID: 15980569 [Indexed for MEDLINE]

3092. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W705-9.

AISMIG--an interactive server-side molecule image generator.

Bohne-Lang A(1), Groch WD, Ranzinger R.

Author information:

(1)German Cancer Research Center Heidelberg, Central Spectroscopy-Molecular

Modeling, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany.

Using a web browser without additional software and generating interactive high

quality and high resolution images of bio-molecules is no longer a problem.

Interactive visualization of 3D molecule structures by Internet browsers normally

is not possible without additional software and the disadvantage of browser-based

structure images (e.g. by a Java applet) is their low resolution. Scientists who

want to generate 3D molecular images with high quality and high resolution (e.g.

for publications or to render a molecule for a poster) therefore require

separately installed software that is often not easy to use. The alternative

concept is an interactive server-side rendering application that can be

interfaced with any web browser. Thus it combines the advantage of the web

application with the high-end rendering of a raytracer. This article addresses

users who want to generate high quality images from molecular structures and do

not have software installed locally for structure visualization. Often people do

not have a structure viewer, such as RasMol or Chime (or even Java) installed

locally but want to visualize a molecule structure interactively. AISMIG (An

Interactive Server-side Molecule Image Generator) is a web service that provides

a visualization of molecule structures in such cases. AISMIG-URL:

http://www.dkfz-heidelberg.de/spec/aismig/.

DOI: 10.1093/nar/gki438

PMCID: PMC1160199

PMID: 15980568 [Indexed for MEDLINE]

3093. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W701-4.

miRU: an automated plant miRNA target prediction server.

Zhang Y(1).

Author information:

(1)Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, OK

73402, USA. yjzhang@noble.org

MicroRNAs (miRNAs) play important roles in gene expression regulation in animals

and plants. Since plant miRNAs recognize their target mRNAs by near-perfect base

pairing, computational sequence similarity search can be used to identify

potential targets. A web-based integrated computing system, miRU, has been

developed for plant miRNA target gene prediction in any plant, if a large number

of sequences are available. Given a mature miRNA sequence from a plant species,

the system thoroughly searches for potential complementary target sites with

mismatches tolerable in miRNA-target recognition. True or false positives are

estimated based on the number and type of mismatches in the target site, and on

the evolutionary conservation of target complementarity in another genome which

can be selected according to miRNA conservation. The output for predicted

targets, ordered by mismatch scores, includes complementary sequences with

mismatches highlighted in colors, original gene sequences and associated

functional annotations. The miRU web server is available at

http://bioinfo3.noble.org/miRU.htm.

DOI: 10.1093/nar/gki383

PMCID: PMC1160144

PMID: 15980567 [Indexed for MEDLINE]

3094. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W696-700.

MicroInspector: a web tool for detection of miRNA binding sites in an RNA

sequence.

Rusinov V(1), Baev V, Minkov IN, Tabler M.

Author information:

(1)Institute of Molecular Biology and Biotechnology, Foundation for Research and

Technology-Hellas, PO Box 1527, GR-71110 Heraklion/Crete, Greece.

Regulation of post-transcriptional gene expression by microRNAs (miRNA) has so

far been validated for only a few mRNA targets. Based on the large number of

miRNA genes and the possibility that one miRNA might influence gene expression of

several targets simultaneously, the quantity of ribo-regulated genes is expected

to be much higher. Here, we describe the web tool MicroInspector that will

analyse a user-defined RNA sequence, which is typically an mRNA or a part of an

mRNA, for the occurrence of binding sites for known and registered miRNAs. The

program allows variation of temperature, the setting of energy values as well as

the selection of different miRNA databases to identify miRNA-binding sites of

different strength. MicroInspector could spot the correct sites for

miRNA-interaction in known target mRNAs. Using other mRNAs, for which such an

interaction has not yet been described, we discovered frequently potential miRNA

binding sites of similar quality, which can now be analysed experimentally. The

MicroInspector program is easy to use and does not require specific computer

skills. The service can be accessed via the MicroInspector web server at

http://www.imbb.forth.gr/microinspector.

DOI: 10.1093/nar/gki364

PMCID: PMC1160125

PMID: 15980566 [Indexed for MEDLINE]

3095. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W690-2.

RibEx: a web server for locating riboswitches and other conserved bacterial

regulatory elements.

Abreu-Goodger C(1), Merino E.

Author information:

(1)Departamento de Microbiología Molecular, Instituto de Biotecnología,

Universidad Nacional Autónoma de México, Cuernavaca, 62210 Morelos, México.

We present RibEx (riboswitch explorer), a web server capable of searching any

sequence for known riboswitches as well as other predicted, but highly conserved,

bacterial regulatory elements. It allows the visual inspection of the identified

motifs in relation to attenuators and open reading frames (ORFs). Any of the

ORF's or regulatory elements' sequence can be obtained with a click and submitted

to NCBI's BLAST. Alternatively, the genome context of all other genes regulated

by the same element can be explored with our genome context tool (GeConT). RibEx

is available at http://www.ibt.unam.mx/biocomputo/ribex.html.

DOI: 10.1093/nar/gki445

PMCID: PMC1160206

PMID: 15980564 [Indexed for MEDLINE]

3096. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W673-6.

GBA server: EST-based digital gene expression profiling.

Wu X(1), Walker MG, Luo J, Wei L.

Author information:

(1)Center for Bioinformatics, National Laboratory of Protein Engineering and

Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing

100871, P. R. China.

Expressed Sequence Tag-based gene expression profiling can be used to discover

functionally associated genes on a large scale. Currently available web servers

and tools focus on finding differentially expressed genes in different samples or

tissues rather than finding co-expressed genes. To fill this gap, we have

developed a web server that implements the GBA (Guilt-by-Association)

co-expression algorithm, which has been successfully used in finding

disease-related genes. We have also annotated UniGene clusters with links to

several important databases such as GO, KEGG, OMIM, Gene, IPI and HomoloGene. The

GBA server can be accessed and downloaded at http://gba.cbi.pku.edu.cn.

DOI: 10.1093/nar/gki480

PMCID: PMC1160240

PMID: 15980560 [Indexed for MEDLINE]

3097. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W659-64.

ASIAN: a web server for inferring a regulatory network framework from gene

expression profiles.

Aburatani S(1), Goto K, Saito S, Toh H, Horimoto K.

Author information:

(1)Laboratory of Biostatistics, Institute of Medical Science, University of

Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan.

The standard workflow in gene expression profile analysis to identify gene

function is the clustering by various metrics and techniques, and the following

analyses, such as sequence analyses of upstream regions. A further challenging

analysis is the inference of a gene regulatory network, and some computational

methods have been intensively developed to deduce the gene regulatory network.

Here, we describe our web server for inferring a framework of regulatory networks

from a large number of gene expression profiles, based on graphical Gaussian

modeling (GGM) in combination with hierarchical clustering

(http://eureka.ims.u-tokyo.ac.jp/asian). GGM is based on a simple mathematical

structure, which is the calculation of the inverse of the correlation coefficient

matrix between variables, and therefore, our server can analyze a wide variety of

data within a reasonable computational time. The server allows users to input the

expression profiles, and it outputs the dendrogram of genes by several

hierarchical clustering techniques, the cluster number estimated by a stopping

rule for hierarchical clustering and the network between the clusters by GGM,

with the respective graphical presentations. Thus, the ASIAN (Automatic System

for Inferring A Network) web server provides an initial basis for inferring

regulatory relationships, in that the clustering serves as the first step toward

identifying the gene function.

DOI: 10.1093/nar/gki446

PMCID: PMC1160207

PMID: 15980557 [Indexed for MEDLINE]

3098. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W650-3.

The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif

search.

Havgaard JH(1), Lyngsø RB, Gorodkin J.

Author information:

(1)Center for Bioinformatics and Division of Genetics, IBHV, The Royal Veterinary

and Agricultural University Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark.

Foldalign is a Sankoff-based algorithm for making structural alignments of RNA

sequences. Here, we present a web server for making pairwise alignments between

two RNA sequences, using the recently updated version of foldalign. The server

can be used to scan two sequences for a common structural RNA motif of limited

size, or the entire sequences can be aligned locally or globally. The web server

offers a graphical interface, which makes it simple to make alignments and

manually browse the results. The web server can be accessed at

http://foldalign.kvl.dk.

DOI: 10.1093/nar/gki473

PMCID: PMC1160234

PMID: 15980555 [Indexed for MEDLINE]

3099. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W65-71.

PAT: a protein analysis toolkit for integrated biocomputing on the web.

Gracy J(1), Chiche L.

Author information:

(1)Centre de Biochimie Structurale, UMR5048 and UMR554 CNRS-INSERM-Université

Montpellier I, Faculté de Pharmacie, 15 avenue Charles Flahault, BP 14491, 34093

Montpellier-Cedex 5, France. jgracy@cbs.cnrs.fr

PAT, for Protein Analysis Toolkit, is an integrated biocomputing server. The main

goal of its design was to facilitate the combination of different processing

tools for complex protein analyses and to simplify the automation of repetitive

tasks. The PAT server provides a standardized web interface to a wide range of

protein analysis tools. It is designed as a streamlined analysis environment that

implements many features which strongly simplify studies dealing with protein

sequences and structures and improve productivity. PAT is able to read and write

data in many bioinformatics formats and to create any desired pipeline by

seamlessly sending the output of a tool to the input of another tool. PAT can

retrieve protein entries from identifier-based queries by using pre-computed

database indexes. Users can easily formulate complex queries combining different

analysis tools with few mouse clicks, or via a dedicated macro language, and a

web session manager provides direct access to any temporary file generated during

the user session. PAT is freely accessible on the Internet at

http://pat.cbs.cnrs.fr.

DOI: 10.1093/nar/gki455

PMCID: PMC1160216

PMID: 15980554 [Indexed for MEDLINE]

3100. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W633-7.

PathwayExplorer: web service for visualizing high-throughput expression data on

biological pathways.

Mlecnik B(1), Scheideler M, Hackl H, Hartler J, Sanchez-Cabo F, Trajanoski Z.

Author information:

(1)Institute for Genomics and Bioinformatics and Christian-Doppler Laboratory for

Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, Graz

8010, Austria.

While generation of high-throughput expression data is becoming routine, the

fast, easy, and systematic presentation and analysis of these data in a

biological context is still an obstacle. To address this need, we have developed

PathwayExplorer, which maps expression profiles of genes or proteins

simultaneously onto major, currently available regulatory, metabolic and cellular

pathways from KEGG, BioCarta and GenMAPP. PathwayExplorer is a

platform-independent web server application with an optional standalone Java

application using a SOAP (simple object access protocol) interface. Mapped

pathways are ranked for the easy selection of the pathway of interest, displaying

all available genes of this pathway with their expression profiles in a

selectable and intuitive color code. Pathway maps produced can be downloaded as

PNG, JPG or as high-resolution vector graphics SVG. The web service is freely

available at https://pathwayexplorer.genome.tugraz.at; the standalone client can

be downloaded at http://genome.tugraz.at.

DOI: 10.1093/nar/gki391

PMCID: PMC1160152

PMID: 15980551 [Indexed for MEDLINE]

3101. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W605-10.

Kinefold web server for RNA/DNA folding path and structure prediction including

pseudoknots and knots.

Xayaphoummine A(1), Bucher T, Isambert H.

Author information:

(1)Laboratoire de Dynamique des Fluides Complexes, CNRS-ULP, Institut de

Physique, 3 rue de l'Université, 67000 Strasbourg, France.

The Kinefold web server provides a web interface for stochastic folding

simulations of nucleic acids on second to minute molecular time scales.

Renaturation or co-transcriptional folding paths are simulated at the level of

helix formation and dissociation in agreement with the seminal experimental

results. Pseudoknots and topologically 'entangled' helices (i.e. knots) are

efficiently predicted taking into account simple geometrical and topological

constraints. To encourage interactivity, simulations launched as immediate jobs

are automatically stopped after a few seconds and return adapted recommendations.

Users can then choose to continue incomplete simulations using the batch queuing

system or go back and modify suggested options in their initial query. Detailed

output provide (i) a series of low free energy structures, (ii) an online

animated folding path and (iii) a programmable trajectory plot focusing on a few

helices of interest to each user. The service can be accessed at

http://kinefold.curie.fr/.

DOI: 10.1093/nar/gki447

PMCID: PMC1160208

PMID: 15980546 [Indexed for MEDLINE]

3102. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W600-4.

RNALOSS: a web server for RNA locally optimal secondary structures.

Clote P(1).

Author information:

(1)Department of Biology, Boston College, Chestnut Hill, MA 02467, USA.

RNAomics, analogous to proteomics, concerns aspects of the secondary and tertiary

structure, folding pathway, kinetics, comparison, function and regulation of all

RNA in a living organism. Given recently discovered roles played by micro RNA,

small interfering RNA, riboswitches, ribozymes, etc., it is important to gain

insight into the folding process of RNA sequences. We describe the web server

RNALOSS, which provides information about the distribution of locally optimal

secondary structures, that possibly form kinetic traps in the folding process.

The tool RNALOSS may be useful in designing RNA sequences which not only have low

folding energy, but whose distribution of locally optimal secondary structures

would suggest rapid and robust folding. Website:

http://clavius.bc.edu/~clotelab/RNALOSS/.

DOI: 10.1093/nar/gki382

PMCID: PMC1160143

PMID: 15980545 [Indexed for MEDLINE]

3103. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W596-9.

GEMS: a web server for biclustering analysis of expression data.

Wu CJ(1), Kasif S.

Author information:

(1)Program in Bioinformatics, Boston University, Boston, MA 02215, USA.

The advent of microarray technology has revolutionized the search for genes that

are differentially expressed across a range of cell types or experimental

conditions. Traditional clustering methods, such as hierarchical clustering, are

often difficult to deploy effectively since genes rarely exhibit similar

expression pattern across a wide range of conditions. Biclustering of gene

expression data (also called co-clustering or two-way clustering) is a

non-trivial but promising methodology for the identification of gene groups that

show a coherent expression profile across a subset of conditions. Thus,

biclustering is a natural methodology as a screen for genes that are functionally

related, participate in the same pathways, affected by the same drug or

pathological condition, or genes that form modules that are potentially

co-regulated by a small group of transcription factors. We have developed a

web-enabled service called GEMS (Gene Expression Mining Server) for biclustering

microarray data. Users may upload expression data and specify a set of criteria.

GEMS then performs bicluster mining based on a Gibbs sampling paradigm. The web

server provides a flexible and an useful platform for the discovery of

co-expressed and potentially co-regulated gene modules. GEMS is an open source

software and is available at http://genomics10.bu.edu/terrence/gems/.

DOI: 10.1093/nar/gki469

PMCID: PMC1160230

PMID: 15980544 [Indexed for MEDLINE]

3104. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W577-81.

DINAMelt web server for nucleic acid melting prediction.

Markham NR(1), Zuker M.

Author information:

(1)Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY

12180-3590, USA.

The DINAMelt web server simulates the melting of one or two single-stranded

nucleic acids in solution. The goal is to predict not just a melting temperature

for a hybridized pair of nucleic acids, but entire equilibrium melting profiles

as a function of temperature. The two molecules are not required to be

complementary, nor must the two strand concentrations be equal. Competition among

different molecular species is automatically taken into account. Calculations

consider not only the heterodimer, but also the two possible homodimers, as well

as the folding of each single-stranded molecule. For each of these five molecular

species, free energies are computed by summing Boltzmann factors over every

possible hybridized or folded state. For temperatures within a user-specified

range, calculations predict species mole fractions together with the free energy,

enthalpy, entropy and heat capacity of the ensemble. Ultraviolet (UV) absorbance

at 260 nm is simulated using published extinction coefficients and computed base

pair probabilities. All results are available as text files and plots are

provided for species concentrations, heat capacity and UV absorbance versus

temperature. This server is connected to an active research program and should

evolve as new theory and software are developed. The server URL is

http://www.bioinfo.rpi.edu/applications/hybrid/.

DOI: 10.1093/nar/gki591

PMCID: PMC1160267

PMID: 15980540 [Indexed for MEDLINE]

3105. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W573-6.

Stitchprofiles.uio.no: analysis of partly melted DNA conformations using stitch

profiles.

Tøstesen E(1), Jerstad GI, Hovig E.

Author information:

(1)The Norwegian Radium Hospital, N-0310 Oslo, Norway.

eivind.tostesen@medisin.uio.no

In this study, we describe a web server that performs computations on DNA

melting, thus predicting the localized separation of the two strands for

sequences provided by the users. The output types are stitch profiles, melting

curves, probability profiles, etc. Stitch profile diagrams visualize the ensemble

of alternative conformations that DNA can adopt with different probabilities. For

example, a stitch profile shows the possible loop openings in terms of their

locations, sizes, probabilities and fluctuations at a given temperature.

Sequences with lengths up to several tens or hundreds of kilobase pairs can be

analysed. The tools are freely available at http://stitchprofiles.uio.no.

DOI: 10.1093/nar/gki424

PMCID: PMC1160185

PMID: 15980539 [Indexed for MEDLINE]

3106. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W570-2.

dnaMATE: a consensus melting temperature prediction server for short DNA

sequences.

Panjkovich A(1), Norambuena T, Melo F.

Author information:

(1)Departamento de Genética Molecular y Microbiología, Facultad de Ciencias

Biológicas, Pontificia Universidad Católica de Chile, Alameda 340, Santiago,

Chile.

An accurate and robust large-scale melting temperature prediction server for

short DNA sequences is dispatched. The server calculates a consensus melting

temperature value using the nearest-neighbor model based on three independent

thermodynamic data tables. The consensus method gives an accurate prediction of

melting temperature, as it has been recently demonstrated in a benchmark

performed using all available experimental data for DNA sequences within the

length range of 16-30 nt. This constitutes the first web server that has been

implemented to perform a large-scale calculation of melting temperatures in real

time (up to 5000 DNA sequences can be submitted in a single run). The expected

accuracy of calculations carried out by this server in the range of 50-600 mM

monovalent salt concentration is that 89% of the melting temperature predictions

will have an error or deviation of <5 degrees C from experimental data. The

server can be freely accessed at http://dna.bio.puc.cl/tm.html. The standalone

executable versions of this software for LINUX, Macintosh and Windows platforms

are also freely available at the same web site. Detailed further information

supporting this server is available at the same web site referenced above.

DOI: 10.1093/nar/gki379

PMCID: PMC1160140

PMID: 15980538 [Indexed for MEDLINE]

3107. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W567-9.

FeatureExtract--extraction of sequence annotation made easy.

Wernersson R(1).

Author information:

(1)Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University

of Denmark, Building 208, DK-2800 Lyngby, Denmark. raz@cbs.dtu.dk

Work on a large number of biological problems benefits tremendously from having

an easy way to access the annotation of DNA sequence features, such as

intron/exon structure, the contents of promoter regions and the location of other

genes in upsteam and downstream regions. For example, taking the placement of

introns within a gene into account can help in a phylogenetic analysis of

homologous genes. Designing experiments for investigating UTR regions using PCR

or DNA microarrays require knowledge of known elements in UTR regions and the

positions and strandness of other genes nearby on the chromosome. A wealth of

such information is already known and documented in databases such as GenBank and

the NCBI Human Genome builds. However, it usually requires significant

bioinformatics skills and intimate knowledge of the data format to access this

information. Presented here is a highly flexible and easy-to-use tool for

extracting feature annotation from GenBank entries. The tool is also useful for

extracting datasets corresponding to a particular feature (e.g. promoters). Most

importantly, the output data format is highly consistent, easy to handle for the

user and easy to parse computationally. The FeatureExtract web server is freely

available for both academic and commercial use at

http://www.cbs.dtu.dk/services/FeatureExtract/.

DOI: 10.1093/nar/gki388

PMCID: PMC1160149

PMID: 15980537 [Indexed for MEDLINE]

3108. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W557-9.

PHYML Online--a web server for fast maximum likelihood-based phylogenetic

inference.

Guindon S(1), Lethiec F, Duroux P, Gascuel O.

Author information:

(1)Bioinformatics Institute and Allan Wilson Centre, University of Auckland

Private Bag 92019, Auckland, New Zealand.

PHYML Online is a web interface to PHYML, a software that implements a fast and

accurate heuristic for estimating maximum likelihood phylogenies from DNA and

protein sequences. This tool provides the user with a number of options, e.g.

nonparametric bootstrap and estimation of various evolutionary parameters, in

order to perform comprehensive phylogenetic analyses on large datasets in

reasonable computing time. The server and its documentation are available at

http://atgc.lirmm.fr/phyml.

DOI: 10.1093/nar/gki352

PMCID: PMC1160113

PMID: 15980534 [Indexed for MEDLINE]

3109. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W544-7.

MuPlex: multi-objective multiplex PCR assay design.

Rachlin J(1), Ding C, Cantor C, Kasif S.

Author information:

(1)Bioinformatics Program, Boston University, Boston, MA 02215, USA.

rachlin@bu.edu

We have developed a web-enabled system called MuPlex that aids researchers in the

design of multiplex PCR assays. Multiplex PCR is a key technology for an endless

list of applications, including detecting infectious microorganisms, whole-genome

sequencing and closure, forensic analysis and for enabling flexible yet low-cost

genotyping. However, the design of a multiplex PCR assays is computationally

challenging because it involves tradeoffs among competing objectives, and

extensive computational analysis is required in order to screen out primer-pair

cross interactions. With MuPlex, users specify a set of DNA sequences along with

primer selection criteria, interaction parameters and the target multiplexing

level. MuPlex designs a set of multiplex PCR assays designed to cover as many of

the input sequences as possible. MuPlex provides multiple solution alternatives

that reveal tradeoffs among competing objectives. MuPlex is uniquely designed for

large-scale multiplex PCR assay design in an automated high-throughput

environment, where high coverage of potentially thousands of single nucleotide

polymorphisms is required. The server is available at

http://genomics14.bu.edu:8080/MuPlex/MuPlex.html.

DOI: 10.1093/nar/gki377

PMCID: PMC1160138

PMID: 15980531 [Indexed for MEDLINE]

3110. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W532-4.

Multiple alignment of genomic sequences using CHAOS, DIALIGN and ABC.

Pöhler D(1), Werner N, Steinkamp R, Morgenstern B.

Author information:

(1)Institute of Microbiology and Genetics, University of Göttingen,

Goldschmidtstr. 1, 37077 Göttingen, Germany.

Comparative analysis of genomic sequences is a powerful approach to discover

functional sites in these sequences. Herein, we present a WWW-based software

system for multiple alignment of genomic sequences. We use the local alignment

tool CHAOS to rapidly identify chains of pairwise similarities. These

similarities are used as anchor points to speed up the DIALIGN multiple-alignment

program. Finally, the visualization tool ABC is used for interactive graphical

representation of the resulting multiple alignments. Our software is available at

Göttingen Bioinformatics Compute Server (GOBICS) at

http://dialign.gobics.de/chaos-dialign-submission.

DOI: 10.1093/nar/gki386

PMCID: PMC1160147

PMID: 15980528 [Indexed for MEDLINE]

3111. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W506-11.

AMOD: a morpholino oligonucleotide selection tool.

Klee EW(1), Shim KJ, Pickart MA, Ekker SC, Ellis LB.

Author information:

(1)Department of Laboratory Medicine and Pathology, University of Minnesota,

Minneapolis, MN, USA.

AMOD is a web-based program that aids in the functional evaluation of nucleotide

sequences through sequence characterization and antisense morpholino

oligonucleotide (target site) selection. Submitted sequences are analyzed by

translation initiation site prediction algorithms and sequence-to-sequence

comparisons; results are used to characterize sequence features required for

morpholino design. Within a defined subsequence, base composition and

homodimerization values are computed for all putative morpholino

oligonucleotides. Using these properties, morpholino candidates are selected and

compared with genomic and transcriptome databases with the goal to identify

target-specific enriched morpholinos. AMOD has been used at the University of

Minnesota to design approximately 200 morpholinos for a functional genomics

screen in zebrafish. The AMOD web server and a tutorial are freely available to

both academic and commercial users at http://www.secretomes.umn.edu/AMOD/.

DOI: 10.1093/nar/gki453

PMCID: PMC1160214

PMID: 15980523 [Indexed for MEDLINE]

3112. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W480-2.

nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide

polymorphisms.

Bao L(1), Zhou M, Cui Y.

Author information:

(1)Department of Molecular Sciences, Center of Genomics and Bioinformatics,

University of Tennessee Health Science Center, 858 Madison Avenue, Memphis, TN

38163, USA.

Nonsynonymous single nucleotide polymorphisms (nsSNPs) are prevalent in genomes

and are closely associated with inherited diseases. To facilitate identifying

disease-associated nsSNPs from a large number of neutral nsSNPs, it is important

to develop computational tools to predict the nsSNP's phenotypic effect

(disease-associated versus neutral). nsSNPAnalyzer, a web-based software

developed for this purpose, extracts structural and evolutionary information from

a query nsSNP and uses a machine learning method called Random Forest to predict

the nsSNP's phenotypic effect. nsSNPAnalyzer server is available at

http://snpanalyzer.utmem.edu/.

DOI: 10.1093/nar/gki372

PMCID: PMC1160133

PMID: 15980516 [Indexed for MEDLINE]

3113. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W468-70.

Integrating protein annotation resources through the Distributed Annotation

System.

Olason PI(1).

Author information:

(1)Center for Biological Sequence, Analysis BioCentrum-DTU, Building 208,

Technical University of Denmark, DK-2800 Lyngby, Denmark. pall@cbs.dtu.dk

Using the Distributed Annotation System (DAS) we have created a protein

annotation resource available at our web page: http://www.cbs.dtu.dk, as a part

of the BioSapiens Network of Excellence EU FP6 project. The DAS protocol allows

us to gather layers of annotation data for a given sequence and thereby gain an

overview of the sequence's features. A user-friendly graphical client has also

been developed (http://www.cbs.dtu.dk/cgi-bin/das), which demonstrates the

possibility of integrating DAS annotation data from multiple sources into a

simple graphical view. The client displays protein feature annotations from the

Center for Biological Sequence Analysis as well as from the BioSapiens reference

UniProt server (http://www.ebi.ac.uk/das-srv/uniprot/das) at the European

Bioinformatics Institute. Other DAS data sources for protein annotation will be

added as they become available.

DOI: 10.1093/nar/gki463

PMCID: PMC1160224

PMID: 15980514 [Indexed for MEDLINE]

3114. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W465-7.

AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined

constraints.

Stanke M(1), Morgenstern B.

Author information:

(1)Universität Göttingen, Institut für Mikrobiologie und Genetik, Abteilung

Bioinformatik, Goldschmidtstrasse 1, 37077 Göttingen, Germany. mstanke@gwdg.de

We present a WWW server for AUGUSTUS, a software for gene prediction in

eukaryotic genomic sequences that is based on a generalized hidden Markov model,

a probabilistic model of a sequence and its gene structure. The web server allows

the user to impose constraints on the predicted gene structure. A constraint can

specify the position of a splice site, a translation initiation site or a stop

codon. Furthermore, it is possible to specify the position of known exons and

intervals that are known to be exonic or intronic sequence. The number of

constraints is arbitrary and constraints can be combined in order to pin down

larger parts of the predicted gene structure. The result then is the most likely

gene structure that complies with all given user constraints, if such a gene

structure exists. The specification of constraints is useful when part of the

gene structure is known, e.g. by expressed sequence tag or protein sequence

alignments, or if the user wants to change the default prediction. The web

interface and the downloadable stand-alone program are available free of charge

at http://augustus.gobics.de/submission.

DOI: 10.1093/nar/gki458

PMCID: PMC1160219

PMID: 15980513 [Indexed for MEDLINE]

3115. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W455-9.

BASys: a web server for automated bacterial genome annotation.

Van Domselaar GH(1), Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P,

Szafron D, Greiner R, Wishart DS.

Author information:

(1)Department of Biological Sciences and Computing Science, University of

Alberta, Edmonton, AB, T6G 2E8, Canada.

BASys (Bacterial Annotation System) is a web server that supports automated,

in-depth annotation of bacterial genomic (chromosomal and plasmid) sequences. It

accepts raw DNA sequence data and an optional list of gene identification

information and provides extensive textual annotation and hyperlinked image

output. BASys uses >30 programs to determine approximately 60 annotation

subfields for each gene, including gene/protein name, GO function, COG function,

possible paralogues and orthologues, molecular weight, isoelectric point, operon

structure, subcellular localization, signal peptides, transmembrane regions,

secondary structure, 3D structure, reactions and pathways. The depth and detail

of a BASys annotation matches or exceeds that found in a standard SwissProt

entry. BASys also generates colorful, clickable and fully zoomable maps of each

query chromosome to permit rapid navigation and detailed visual analysis of all

resulting gene annotations. The textual annotations and images that are provided

by BASys can be generated in approximately 24 h for an average bacterial

chromosome (5 Mb). BASys annotations may be viewed and downloaded anonymously or

through a password protected access system. The BASys server and databases can

also be downloaded and run locally. BASys is accessible at

http://wishart.biology.ualberta.ca/basys.

DOI: 10.1093/nar/gki593

PMCID: PMC1160269

PMID: 15980511 [Indexed for MEDLINE]

3116. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W451-4.

GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.

Besemer J(1), Borodovsky M.

Author information:

(1)School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA.

The task of gene identification frequently confronting researchers working with

both novel and well studied genomes can be conveniently and reliably solved with

the help of the GeneMark web software (http://opal.biology.gatech.edu/GeneMark/).

The website provides interfaces to the GeneMark family of programs designed and

tuned for gene prediction in prokaryotic, eukaryotic and viral genomic sequences.

Currently, the server allows the analysis of nearly 200 prokaryotic and >10

eukaryotic genomes using species-specific versions of the software and

pre-computed gene models. In addition, genes in prokaryotic sequences from novel

genomes can be identified using models derived on the spot upon sequence

submission, either by a relatively simple heuristic approach or by the

full-fledged self-training program GeneMarkS. A database of reannotations of

>1000 viral genomes by the GeneMarkS program is also available from the web site.

The GeneMark website is frequently updated to provide the latest versions of the

software and gene models.

DOI: 10.1093/nar/gki487

PMCID: PMC1160247

PMID: 15980510 [Indexed for MEDLINE]

3117. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W447-50.

CONREAL web server: identification and visualization of conserved transcription

factor binding sites.

Berezikov E(1), Guryev V, Cuppen E.

Author information:

(1)Hubrecht Laboratory, Uppsalalaan 8, 3584 CT, Utrecht, The Netherlands.

berezikov@niob.knaw.nl

The use of orthologous sequences and phylogenetic footprinting approaches have

become popular for the recognition of conserved and potentially functional

sequences. Several algorithms have been developed for the identification of

conserved transcription factor binding sites (TFBSs), which are characterized by

their relatively short and degenerative recognition sequences. The CONREAL

(conserved regulatory elements anchored alignment) web server provides a

versatile interface to CONREAL-, LAGAN-, BLASTZ- and AVID-based predictions of

conserved TFBSs in orthologous promoters. Comparative analysis using different

algorithms can be started by keyword without any prior sequence retrieval. The

interface is available at http://conreal.niob.knaw.nl.

DOI: 10.1093/nar/gki378

PMCID: PMC1160139

PMID: 15980509 [Indexed for MEDLINE]

3118. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W44-9.

RPBS: a web resource for structural bioinformatics.

Alland C(1), Moreews F, Boens D, Carpentier M, Chiusa S, Lonquety M, Renault N,

Wong Y, Cantalloube H, Chomilier J, Hochez J, Pothier J, Villoutreix BO, Zagury

JF, Tufféry P.

Author information:

(1)EBGM, INSERM U726, Université Paris 7, France.

RPBS (Ressource Parisienne en Bioinformatique Structurale) is a resource

dedicated primarily to structural bioinformatics. It is the result of a joint

effort by several teams to set up an interface that offers original and powerful

methods in the field. As an illustration, we focus here on three such methods

uniquely available at RPBS: AUTOMAT for sequence databank scanning, YAKUSA for

structure databank scanning and WLOOP for homology loop modelling. The RPBS

server can be accessed at http://bioserv.rpbs.jussieu.fr/ and the specific

services at http://bioserv.rpbs.jussieu.fr/SpecificServices.html.

DOI: 10.1093/nar/gki477

PMCID: PMC1160237

PMID: 15980507 [Indexed for MEDLINE]

3119. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W423-6.

PromoterPlot: a graphical display of promoter similarities by pattern

recognition.

Di Cara A(1), Schmidt K, Hemmings BA, Oakeley EJ.

Author information:

(1)Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66,

CH-4058 Basel, Switzerland.

PromoterPlot (http://promoterplot.fmi.ch) is a web-based tool for simplifying the

display and processing of transcription factor searches using either the

commercial or free TransFac distributions. The input sequence is a TransFac

search (public version) or FASTA/Affymetrix IDs (local install). It uses an

intuitive pattern recognition algorithm for finding similarities between groups

of promoters by dividing transcription factor predictions into conserved triplet

models. To minimize the number of false-positive models, it can optionally

exclude factors that are known to be unexpressed or inactive in the cells being

studied based on microarray or proteomic expression data. The program will also

estimate the likelihood of finding a pattern by chance based on the frequency

observed in a control set of mammalian promoters we obtained from Genomatix. The

results are stored as an interactive SVG web page on our server.

DOI: 10.1093/nar/gki413

PMCID: PMC1160174

PMID: 15980503 [Indexed for MEDLINE]

3120. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W412-6.

WordSpy: identifying transcription factor binding motifs by building a dictionary

and learning a grammar.

Wang G(1), Yu T, Zhang W.

Author information:

(1)Department of Computer Science and Engineering, Washington University in Saint

Louis, Saint Louis, MO 63130, USA.

Transcription factor (TF) binding sites or motifs (TFBMs) are functional

cis-regulatory DNA sequences that play an essential role in gene transcriptional

regulation. Although many experimental and computational methods have been

developed, finding TFBMs remains a challenging problem. We propose and develop a

novel dictionary based motif finding algorithm, which we call WordSpy. One

significant feature of WordSpy is the combination of a word counting method and a

statistical model which consists of a dictionary of motifs and a grammar

specifying their usage. The algorithm is suitable for genome-wide motif finding;

it is capable of discovering hundreds of motifs from a large set of promoters in

a single run. We further enhance WordSpy by applying gene expression information

to separate true TFBMs from spurious ones, and by incorporating negative

sequences to identify discriminative motifs. In addition, we also use randomly

selected promoters from the genome to evaluate the significance of the discovered

motifs. The output from WordSpy consists of an ordered list of putative motifs

and a set of regulatory sequences with motif binding sites highlighted. The web

server of WordSpy is available at http://cic.cs.wustl.edu/wordspy.

DOI: 10.1093/nar/gki492

PMCID: PMC1160252

PMID: 15980501 [Indexed for MEDLINE]

3121. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W382-8.

The FoldX web server: an online force field.

Schymkowitz J(1), Borg J, Stricher F, Nys R, Rousseau F, Serrano L.

Author information:

(1)Switch Laboratory, Flanders Interuniversity Institute for Biotechnology (VIB),

Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium.

joost.schymkowitz@skynet.be

FoldX is an empirical force field that was developed for the rapid evaluation of

the effect of mutations on the stability, folding and dynamics of proteins and

nucleic acids. The core functionality of FoldX, namely the calculation of the

free energy of a macromolecule based on its high-resolution 3D structure, is now

publicly available through a web server at http://foldx.embl.de/. The current

release allows the calculation of the stability of a protein, calculation of the

positions of the protons and the prediction of water bridges, prediction of metal

binding sites and the analysis of the free energy of complex formation. Alanine

scanning, the systematic truncation of side chains to alanine, is also included.

In addition, some reporting functions have been added, and it is now possible to

print both the atomic interaction networks that constitute the protein, print the

structural and energetic details of the interactions per atom or per residue, as

well as generate a general quality report of the pdb structure. This core

functionality will be further extended as more FoldX applications are developed.

DOI: 10.1093/nar/gki387

PMCID: PMC1160148

PMID: 15980494 [Indexed for MEDLINE]

3122. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W368-71.

H++: a server for estimating pKas and adding missing hydrogens to macromolecules.

Gordon JC(1), Myers JB, Folta T, Shoja V, Heath LS, Onufriev A.

Author information:

(1)Department of Computer Science, Virginia Tech Blacksburg, VA 24061, USA.

The structure and function of macromolecules depend critically on the ionization

(protonation) states of their acidic and basic groups. A number of existing

practical methods predict protonation equilibrium pK constants of macromolecules

based upon their atomic resolution Protein Data Bank (PDB) structures; the

calculations are often performed within the framework of the continuum

electrostatics model. Unfortunately, these methodologies are complex, involve

multiple steps and require considerable investment of effort. Our web server

http://biophysics.cs.vt.edu/H++ provides access to a tool that automates this

process, allowing both experts and novices to quickly obtain estimates of pKs as

well as other related characteristics of biomolecules such as isoelectric points,

titration curves and energies of protonation microstates. Protons are added to

the input structure according to the calculated ionization states of its

titratable groups at the user-specified pH; the output is in the PQR (PDB +

charges + radii) format. In addition, corresponding coordinate and topology files

are generated in the format supported by the molecular modeling package AMBER.

The server is intended for a broad community of biochemists, molecular modelers,

structural biologists and drug designers; it can also be used as an educational

tool in biochemistry courses.

DOI: 10.1093/nar/gki464

PMCID: PMC1160225

PMID: 15980491 [Indexed for MEDLINE]

3123. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W36-8.

Protein structure prediction servers at University College London.

Bryson K(1), McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT.

Author information:

(1)Department of Computer Science, University College London Gower Street, London

WC1E 6BT, UK.

A number of state-of-the-art protein structure prediction servers have been

developed by researchers working in the Bioinformatics Unit at University College

London. The popular PSIPRED server allows users to perform secondary structure

prediction, transmembrane topology prediction and protein fold recognition. More

recent servers include DISOPRED for the prediction of protein dynamic disorder

and DomPred for domain boundary prediction. These servers are available from our

software home page at http://bioinf.cs.ucl.ac.uk/software.html.

DOI: 10.1093/nar/gki410

PMCID: PMC1160171

PMID: 15980489 [Indexed for MEDLINE]

3124. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W358-62.

MovieMaker: a web server for rapid rendering of protein motions and interactions.

Maiti R(1), Van Domselaar GH, Wishart DS.

Author information:

(1)Department of Computing Science, University of Alberta, Edmonton, AB, Canada

T6G 2E8.

MovieMaker is a web server that allows short ( approximately 10 s), downloadable

movies of protein motions to be generated. It accepts PDB files or PDB accession

numbers as input and automatically calculates, renders and merges the necessary

image files to create colourful animations covering a wide range of protein

motions and other dynamic processes. Users have the option of animating (i)

simple rotation, (ii) morphing between two end-state conformers, (iii)

short-scale, picosecond vibrations, (iv) ligand docking, (v) protein

oligomerization, (vi) mid-scale nanosecond (ensemble) motions and (vii) protein

folding/unfolding. MovieMaker does not perform molecular dynamics calculations.

Instead it is an animation tool that uses a sophisticated superpositioning

algorithm in conjunction with Cartesian coordinate interpolation to rapidly and

automatically calculate the intermediate structures needed for many of its

animations. Users have extensive control over the rendering style, structure

colour, animation quality, background and other image features. MovieMaker is

intended to be a general-purpose server that allows both experts and non-experts

to easily generate useful, informative protein animations for educational and

illustrative purposes. MovieMaker is accessible at

http://wishart.biology.ualberta.ca/moviemaker.

DOI: 10.1093/nar/gki485

PMCID: PMC1160245

PMID: 15980488 [Indexed for MEDLINE]

3125. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W342-6.

POPSCOMP: an automated interaction analysis of biomolecular complexes.

Kleinjung J(1), Fraternali F.

Author information:

(1)Bioinformatics Unit, Faculty of Sciences, Vrije Universiteit De Boelelaan

1081A, 1081HV, Amsterdam, The Netherlands.

Large-scale analysis of biomolecular complexes reveals the functional network

within the cell. Computational methods are required to extract the essential

information from the available data. The POPSCOMP server is designed to calculate

the interaction surface between all components of a given complex structure

consisting of proteins, DNA or RNA molecules. The server returns matrices and

graphs of surface area burial that can be used to automatically annotate

components and residues that are involved in complex formation, to pinpoint

conformational changes and to estimate molecular interaction energies. The

analysis can be performed on a per-atom level or alternatively on a per-residue

level for low-resolution structures. Here, we present an analysis of ribosomal

structures in complex with various antibiotics to exemplify the potential and

limitations of automated complex analysis. The POPSCOMP server is accessible at

http://ibivu.cs.vu.nl/programs/popscompwww/.

DOI: 10.1093/nar/gki369

PMCID: PMC1160130

PMID: 15980485 [Indexed for MEDLINE]

3126. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W326-30.

Metabolic PathFinding: inferring relevant pathways in biochemical networks.

Croes D(1), Couche F, Wodak SJ, van Helden J.

Author information:

(1)SCMBB, Université Libre de Bruxelles, Campus Plaine, CP 263, Boulevard du

Triomphe, B-1050 Bruxelles, Belgium.

Our knowledge of metabolism can be represented as a network comprising several

thousands of nodes (compounds and reactions). Several groups applied graph theory

to analyse the topological properties of this network and to infer metabolic

pathways by path finding. This is, however, not straightforward, with a major

problem caused by traversing irrelevant shortcuts through highly connected nodes,

which correspond to pool metabolites and co-factors (e.g. H2O, NADP and H+). In

this study, we present a web server implementing two simple approaches, which

circumvent this problem, thereby improving the relevance of the inferred

pathways. In the simplest approach, the shortest path is computed, while

filtering out the selection of highly connected compounds. In the second

approach, the shortest path is computed on the weighted metabolic graph where

each compound is assigned a weight equal to its connectivity in the network. This

approach significantly increases the accuracy of the inferred pathways, enabling

the correct inference of relatively long pathways (e.g. with as many as eight

intermediate reactions). Available options include the calculation of the

k-shortest paths between two specified seed nodes (either compounds or

reactions). Multiple requests can be submitted in a queue. Results are returned

by email, in textual as well as graphical formats (available in

http://www.scmbb.ulb.ac.be/pathfinding/).

DOI: 10.1093/nar/gki437

PMCID: PMC1160198

PMID: 15980483 [Indexed for MEDLINE]

3127. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W324-5.

BIOVERSE: enhancements to the framework for structural, functional and contextual

modeling of proteins and proteomes.

McDermott J(1), Guerquin M, Frazier Z, Chang AN, Samudrala R.

Author information:

(1)Department of Microbiology, University of Washington, Seattle, WA, USA.

We have made a number of enhancements to the previously described Bioverse web

server and computational biology framework

(http://bioverse.compbio.washington.edu). In this update, we provide an overview

of the new features available that include: (i) expansion of the number of

organisms represented in the Bioverse and addition of new data sources and novel

prediction techniques not available elsewhere, including network-based

annotation; (ii) reengineering the database backend and supporting code resulting

in significant speed, search and ease-of use improvements; and (iii) creation of

a stateful and dynamic web application frontend to improve interface speed and

usability. Integrated Java-based applications also allow dynamic visualization of

real and predicted protein interaction networks.

DOI: 10.1093/nar/gki401

PMCID: PMC1160162

PMID: 15980482 [Indexed for MEDLINE]

3128. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W306-10.

I-Mutant2.0: predicting stability changes upon mutation from the protein sequence

or structure.

Capriotti E(1), Fariselli P, Casadio R.

Author information:

(1)Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna

via Irnerio 42, 40126 Bologna, Italy.

I-Mutant2.0 is a support vector machine (SVM)-based tool for the automatic

prediction of protein stability changes upon single point mutations. I-Mutant2.0

predictions are performed starting either from the protein structure or, more

importantly, from the protein sequence. This latter task, to the best of our

knowledge, is exploited for the first time. The method was trained and tested on

a data set derived from ProTherm, which is presently the most comprehensive

available database of thermodynamic experimental data of free energy changes of

protein stability upon mutation under different conditions. I-Mutant2.0 can be

used both as a classifier for predicting the sign of the protein stability change

upon mutation and as a regression estimator for predicting the related

DeltaDeltaG values. Acting as a classifier, I-Mutant2.0 correctly predicts (with

a cross-validation procedure) 80% or 77% of the data set, depending on the usage

of structural or sequence information, respectively. When predicting DeltaDeltaG

values associated with mutations, the correlation of predicted with

expected/experimental values is 0.71 (with a standard error of 1.30 kcal/mol) and

0.62 (with a standard error of 1.45 kcal/mol) when structural or sequence

information are respectively adopted. Our web interface allows the selection of a

predictive mode that depends on the availability of the protein structure and/or

sequence. In this latter case, the web server requires only pasting of a protein

sequence in a raw format. We therefore introduce I-Mutant2.0 as a unique and

valuable helper for protein design, even when the protein structure is not yet

known with atomic resolution.AVAILABILITY:

http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi.

DOI: 10.1093/nar/gki375

PMCID: PMC1160136

PMID: 15980478 [Indexed for MEDLINE]

3129. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W303-5.

SRide: a server for identifying stabilizing residues in proteins.

Magyar C(1), Gromiha MM, Pujadas G, Tusnády GE, Simon I.

Author information:

(1)Institute of Enzymology, Biological Research Center, Hungarian Academy of

Sciences H-1518 Budapest, PO Box 7, Hungary.

Residues expected to play key roles in the stabilization of proteins [stabilizing

residues (SRs)] are selected by combining several methods based mainly on the

interactions of a given residue with its spatial, rather than its sequential

neighborhood and by considering the evolutionary conservation of the residues. A

residue is selected as a stabilizing residue if it has high surrounding

hydrophobicity, high long-range order, high conservation score and if it belongs

to a stabilization center. The definition of all these parameters and the

thresholds used to identify the SRs are discussed in detail. The algorithm for

identifying SRs was originally developed for TIM-barrel proteins [M. M. Gromiha,

G. Pujadas, C. Magyar, S. Selvaraj, and I. Simon (2004), Proteins, 55, 316-329]

and is now generalized for all proteins of known 3D structure. SRs could be

applied in protein engineering and homology modeling and could also help to

explain certain folds with significant stability. The SRide server is located at

http://sride.enzim.hu.

DOI: 10.1093/nar/gki409

PMCID: PMC1160170

PMID: 15980477 [Indexed for MEDLINE]

3130. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W29-35.

The Diamond STING server.

Neshich G(1), Borro LC, Higa RH, Kuser PR, Yamagishi ME, Franco EH, Krauchenco

JN, Fileto R, Ribeiro AA, Bezerra GB, Velludo TM, Jimenez TS, Furukawa N, Teshima

H, Kitajima K, Bava A, Sarai A, Togawa RC, Mancini AL.

Author information:

(1)Núcleo de Bioinformática Estrutural, Embrapa/Informática Agropecuária

Campinas, Brazil. neshich@cnptia.embrapa.br

Diamond STING is a new version of the STING suite of programs for a comprehensive

analysis of a relationship between protein sequence, structure, function and

stability. We have added a number of new functionalities by both providing more

structure parameters to the STING Database and by improving/expanding the

interface for enhanced data handling. The integration among the STING components

has also been improved. A new key feature is the ability of the STING server to

handle local files containing protein structures (either modeled or not yet

deposited to the Protein Data Bank) so that they can be used by the principal

STING components: (Java)Protein Dossier ((J)PD) and STING Report. The current

capabilities of the new STING version and a couple of biologically relevant

applications are described here. We have provided an example where Diamond STING

identifies the active site amino acids and folding essential amino acids (both

previously determined by experiments) by filtering out all but those residues by

selecting the numerical values/ranges for a set of corresponding parameters. This

is the fundamental step toward a more interesting endeavor-the prediction of such

residues. Diamond STING is freely accessible at http://sms.cbi.cnptia.embrapa.br

and http://trantor.bioc.columbia.edu/SMS.

DOI: 10.1093/nar/gki397

PMCID: PMC1160158

PMID: 15980473 [Indexed for MEDLINE]

3131. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W284-8.

FFAS03: a server for profile--profile sequence alignments.

Jaroszewski L(1), Rychlewski L, Li Z, Li W, Godzik A.

Author information:

(1)Bioinformatics Program, The Burnham Institute, 10901 N. Torrey Pines Road, La

Jolla, CA 92037, USA.

The FFAS03 server provides a web interface to the third generation of the

profile-profile alignment and fold-recognition algorithm of fold and function

assignment system (FFAS) [L. Rychlewski, L. Jaroszewski, W. Li and A. Godzik

(2000), Protein Sci., 9, 232-241]. Profile-profile algorithms use information

present in sequences of homologous proteins to amplify the patterns defining the

family. As a result, they enable detection of remote homologies beyond the reach

of other methods. FFAS, initially developed in 2000, is consistently one of the

best ranked fold prediction methods in the CAFASP and LiveBench competitions. It

is also used by several fold-recognition consensus methods and meta-servers. The

FFAS03 server accepts a user supplied protein sequence and automatically

generates a profile, which is then compared with several sets of sequence

profiles of proteins from PDB, COG, PFAM and SCOP. The profile databases used by

the server are automatically updated with the latest structural and sequence

information. The server provides access to the alignment analysis, multiple

alignment, and comparative modeling tools. Access to the server is open for both

academic and commercial researchers. The FFAS03 server is available at

http://ffas.burnham.org.

DOI: 10.1093/nar/gki418

PMCID: PMC1160179

PMID: 15980471 [Indexed for MEDLINE]

3132. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W274-6.

SCANMOT: searching for similar sequences using a simultaneous scan of multiple

sequence motifs.

Chakrabarti S(1), Anand AP, Bhardwaj N, Pugalenthi G, Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences (NCBS), Bangalore 560065, India.

Establishment of similarities between proteins is very important for the study of

the relationship between sequence, structure and function and for the analysis of

evolutionary relationships. Motif-based search methods play a crucial role in

establishing the connections between proteins that are particularly useful for

distant relationships. This paper reports SCANMOT, a web-based server that

searches for similarities between proteins by simultaneous matching of multiple

motifs. SCANMOT searches for similar sequences in entire sequence databases using

multiple conserved regions and utilizes inter-motif spacing as restraints. The

SCANMOT server is available via

http://www.ncbs.res.in/~faculty/mini/scanmot/scanmot.html.

DOI: 10.1093/nar/gki493

PMCID: PMC1160253

PMID: 15980468 [Indexed for MEDLINE]

3133. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W262-6.

PatMatch: a program for finding patterns in peptide and nucleotide sequences.

Yan T(1), Yoo D, Berardini TZ, Mueller LA, Weems DC, Weng S, Cherry JM, Rhee SY.

Author information:

(1)Department of Plant Biology, Carnegie Institution of Washington, 260 Panama

Street, Stanford, CA 94305, USA.

Here, we present PatMatch, an efficient, web-based pattern-matching program that

enables searches for short nucleotide or peptide sequences such as cis-elements

in nucleotide sequences or small domains and motifs in protein sequences. The

program can be used to find matches to a user-specified sequence pattern that can

be described using ambiguous sequence codes and a powerful and flexible pattern

syntax based on regular expressions. A recent upgrade has improved performance

and now supports both mismatches and wildcards in a single pattern. This

enhancement has been achieved by replacing the previous searching algorithm,

scan\_for\_matches [D'Souza et al. (1997), Trends in Genetics, 13, 497-498], with

nondeterministic-reverse grep (NR-grep), a general pattern matching tool that

allows for approximate string matching [Navarro (2001), Software Practice and

Experience, 31, 1265-1312]. We have tailored NR-grep to be used for DNA and

protein searches with PatMatch. The stand-alone version of the software can be

adapted for use with any sequence dataset and is available for download at The

Arabidopsis Information Resource (TAIR) at

ftp://ftp.arabidopsis.org/home/tair/Software/Patmatch/. The PatMatch server is

available on the web at

http://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl for searching

Arabidopsis thaliana sequences.

DOI: 10.1093/nar/gki368

PMCID: PMC1160129

PMID: 15980466 [Indexed for MEDLINE]

3134. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W255-61.

QuasiMotiFinder: protein annotation by searching for evolutionarily conserved

motif-like patterns.

Gutman R(1), Berezin C, Wollman R, Rosenberg Y, Ben-Tal N.

Author information:

(1)Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel

Aviv University, Ramat Aviv 69978, Israel.

Sequence signature databases such as PROSITE, which include amino acid segments

that are indicative of a protein's function, are useful for protein annotation.

Lamentably, the annotation is not always accurate. A signature may be falsely

detected in a protein that does not carry out the associated function (false

positive prediction, FP) or may be overlooked in a protein that does carry out

the function (false negative prediction, FN). A new approach has emerged in which

a signature is replaced with a sequence profile, calculated based on multiple

sequence alignment (MSA) of homologous proteins that share the same function.

This approach, which is superior to the simple pattern search, essentially

searches with the sequence of the query protein against an MSA library. We

suggest here an alternative approach, implemented in the QuasiMotiFinder web

server (http://quasimotifinder.tau.ac.il/), which is based on a search with an

MSA of homologous query proteins against the original PROSITE signatures. The

explicit use of the average evolutionary conservation of the signature in the

query proteins significantly reduces the rate of FP prediction compared with the

simple pattern search. QuasiMotiFinder also has a reduced rate of FN prediction

compared with simple pattern searches, since the traditional search for precise

signatures has been replaced by a permissive search for signature-like patterns

that are physicochemically similar to known signatures. Overall, QuasiMotiFinder

and the profile search are comparable to each other in terms of performance. They

are also complementary to each other in that signatures that are falsely detected

in (or overlooked by) one may be correctly detected by the other.

DOI: 10.1093/nar/gki496

PMCID: PMC1160256

PMID: 15980465 [Indexed for MEDLINE]

3135. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W244-8.

The HHpred interactive server for protein homology detection and structure

prediction.

Söding J(1), Biegert A, Lupas AN.

Author information:

(1)Department of Protein Evolution, Max-Planck-Institute for Developmental

Biology Spemannstrasse 35, 72076 Tübingen, Germany.

johannes.soeding@tuebingen.mpg.de

HHpred is a fast server for remote protein homology detection and structure

prediction and is the first to implement pairwise comparison of profile hidden

Markov models (HMMs). It allows to search a wide choice of databases, such as the

PDB, SCOP, Pfam, SMART, COGs and CDD. It accepts a single query sequence or a

multiple alignment as input. Within only a few minutes it returns the search

results in a user-friendly format similar to that of PSI-BLAST. Search options

include local or global alignment and scoring secondary structure similarity.

HHpred can produce pairwise query-template alignments, multiple alignments of the

query with a set of templates selected from the search results, as well as 3D

structural models that are calculated by the MODELLER software from these

alignments. A detailed help facility is available. As a demonstration, we analyze

the sequence of SpoVT, a transcriptional regulator from Bacillus subtilis. HHpred

can be accessed at http://protevo.eb.tuebingen.mpg.de/hhpred.

DOI: 10.1093/nar/gki408

PMCID: PMC1160169

PMID: 15980461 [Indexed for MEDLINE]

3136. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W239-43.

REPPER--repeats and their periodicities in fibrous proteins.

Gruber M(1), Söding J, Lupas AN.

Author information:

(1)Max-Planck Institute for Developmental Biology Spemannstr. 35, 72076 Tübingen,

Germany. markus.gruber@tuebingen.mpg.de

REPPER (REPeats and their PERiodicities) is an integrated server that detects and

analyzes regions with short gapless repeats in protein sequences or alignments.

It finds periodicities by Fourier Transform (FTwin) and internal similarity

analysis (REPwin). FTwin assigns numerical values to amino acids that reflect

certain properties, for instance hydrophobicity, and gives information on

corresponding periodicities. REPwin uses self-alignments and displays repeats

that reveal significant internal similarities. Both programs use a sliding window

to ensure that different periodic regions within the same protein are detected

independently. FTwin and REPwin are complemented by secondary structure

prediction (PSIPRED) and coiled coil prediction (COILS), making the server a

versatile analysis tool for sequences of fibrous proteins. REPPER is available at

http://protevo.eb.tuebingen.mpg.de/repper.

DOI: 10.1093/nar/gki405

PMCID: PMC1160166

PMID: 15980460 [Indexed for MEDLINE]

3137. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W230-2.

DiANNA: a web server for disulfide connectivity prediction.

Ferrè F(1), Clote P.

Author information:

(1)Department of Biology, Boston College, Chestnut Hill, MA 02467, USA.

Correctly predicting the disulfide bond topology in a protein is of crucial

importance for the understanding of protein function and can be of great help for

tertiary prediction methods. The web server

http://clavius.bc.edu/~clotelab/DiANNA/ outputs the disulfide connectivity

prediction given input of a protein sequence. The following procedure is

performed. First, PSIPRED is run to predict the protein's secondary structure,

then PSIBLAST is run against the non-redundant SwissProt to obtain a multiple

alignment of the input sequence. The predicted secondary structure and the

profile arising from this alignment are used in the training phase of our neural

network. Next, cysteine oxidation state is predicted, then each pair of cysteines

in the protein sequence is assigned a likelihood of forming a disulfide

bond--this is performed by means of a novel architecture (diresidue neural

network). Finally, Rothberg's implementation of Gabow's maximum weighted matching

algorithm is applied to diresidue neural network scores in order to produce the

final connectivity prediction. Our novel neural network-based approach achieves

results that are comparable and in some cases better than the current

state-of-the-art methods.

DOI: 10.1093/nar/gki412

PMCID: PMC1160173

PMID: 15980459 [Indexed for MEDLINE]

3138. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W226-9.

KinasePhos: a web tool for identifying protein kinase-specific phosphorylation

sites.

Huang HD(1), Lee TY, Tzeng SW, Horng JT.

Author information:

(1)Department of Biological Science and Technology, Institute of Bioinformatics,

National Chiao Tung University, Hsin-Chu 300, Taiwan.

KinasePhos is a novel web server for computationally identifying catalytic

kinase-specific phosphorylation sites. The known phosphorylation sites from

public domain data sources are categorized by their annotated protein kinases.

Based on the profile hidden Markov model, computational models are learned from

the kinase-specific groups of the phosphorylation sites. After evaluating the

learned models, the model with highest accuracy was selected from each

kinase-specific group, for use in a web-based prediction tool for identifying

protein phosphorylation sites. Therefore, this work developed a kinase-specific

phosphorylation site prediction tool with both high sensitivity and specificity.

The prediction tool is freely available at http://KinasePhos.mbc.nctu.edu.tw/.

DOI: 10.1093/nar/gki471

PMCID: PMC1160232

PMID: 15980458 [Indexed for MEDLINE]

3139. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W198-201.

TRAMPLE: the transmembrane protein labelling environment.

Fariselli P(1), Finelli M, Rossi I, Amico M, Zauli A, Martelli PL, Casadio R.

Author information:

(1)Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna,

via Irnerio 42, I-40126, Bologna, Italy.

TRAMPLE (http://gpcr.biocomp.unibo.it/biodec/) is a web application server

dedicated to the detection and the annotation of transmembrane protein sequences.

TRAMPLE includes different state-of-the-art algorithms for the prediction of

signal peptides, transmembrane segments (both beta-strands and alpha-helices),

secondary structure and fast fold recognition. TRAMPLE also includes a complete

content management system to manage the results of the predictions. Each user of

the server has his/her own workplace, where the data can be stored, organized,

accessed and annotated with documents through a simple web-based interface. In

this manner, TRAMPLE significantly improves usability with respect to other more

traditional web servers.

DOI: 10.1093/nar/gki440

PMCID: PMC1160201

PMID: 15980454 [Indexed for MEDLINE]

3140. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W193-7.

Web-based toolkits for topology prediction of transmembrane helical proteins,

fold recognition, structure and binding scoring, folding-kinetics analysis and

comparative analysis of domain combinations.

Zhou H(1), Zhang C, Liu S, Zhou Y.

Author information:

(1)Department of Physiology & Biophysics, Howard Hughes Medical Institute Center

for Single Molecule Biophysics, State University of New York at Buffalo, 124

Sherman Hall, Buffalo, NY 14214, USA.

We have developed the following web servers for protein structural modeling and

analysis at http://theory.med.buffalo.edu: THUMBUP, UMDHMM(TMHP) and TUPS,

predictors of transmembrane helical protein topology based on a

mean-burial-propensity scale of amino acid residues (THUMBUP), hidden Markov

model (UMDHMM(TMHP)) and their combinations (TUPS); SPARKS 2.0 and SP3, two

profile-profile alignment methods, that match input query sequence(s) to

structural templates by integrating sequence profile with knowledge-based

structural score (SPARKS 2.0) and structure-derived profile (SP3); DFIRE, a

knowledge-based potential for scoring free energy of monomers (DMONOMER), loop

conformations (DLOOP), mutant stability (DMUTANT) and binding affinity of

protein-protein/peptide/DNA complexes (DCOMPLEX & DDNA); TCD, a program for

protein-folding rate and transition-state analysis of small globular proteins;

and DOGMA, a web-server that allows comparative analysis of domain combinations

between plant and other 55 organisms. These servers provide tools for prediction

and/or analysis of proteins on the secondary structure, tertiary structure and

interaction levels, respectively.

DOI: 10.1093/nar/gki360

PMCID: PMC1160121

PMID: 15980453 [Indexed for MEDLINE]

3141. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W188-92.

TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel

proteins.

Garrow AG(1), Agnew A, Westhead DR.

Author information:

(1)School of Biochemistry and Microbiology, University of Leeds, Leeds LS2 9JT,

UK.

TMB-Hunt is a program that uses a modified k-nearest neighbour (k-NN) algorithm

to classify protein sequences as transmembrane beta-barrel (TMB) or non-TMB on

the basis of whole sequence amino acid composition. By including differentially

weighted amino acids, evolutionary information and by calibrating the scoring, a

discrimination accuracy of 92.5% was achieved, as tested using a rigorous

cross-validation procedure. The TMB-Hunt web server, available at

www.bioinformatics.leeds.ac.uk/betaBarrel, allows screening of up to 10,000

sequences in a single query and provides results and key statistics in a simple

colour coded format.

DOI: 10.1093/nar/gki384

PMCID: PMC1160145

PMID: 15980452 [Indexed for MEDLINE]

3142. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W184-7.

GPS: a comprehensive www server for phosphorylation sites prediction.

Xue Y(1), Zhou F, Zhu M, Ahmed K, Chen G, Yao X.

Author information:

(1)School of Life Science, University of Science and Technology of China, Hefei,

Anhui 230027, PR China.

Protein phosphorylation plays a fundamental role in most of the cellular

regulatory pathways. Experimental identification of protein kinases' (PKs)

substrates with their phosphorylation sites is labor-intensive and often limited

by the availability and optimization of enzymatic reactions. Recently,

large-scale analysis of the phosphoproteome by the mass spectrometry (MS) has

become a popular approach. But experimentally, it is still difficult to

distinguish the kinase-specific sites on the substrates. In this regard, the in

silico prediction of phosphorylation sites with their specific kinases using

protein's primary sequences may provide guidelines for further experimental

consideration and interpretation of MS phosphoproteomic data. A variety of such

tools exists over the Internet and provides the predictions for at most 30 PK

subfamilies. We downloaded the verified phosphorylation sites from the public

databases and curated the literature extensively for recently found

phosphorylation sites. With the hypothesis that PKs in the same subfamily share

similar consensus sequences/motifs/functional patterns on substrates, we

clustered the 216 unique PKs in 71 PK groups, according to the BLAST results and

protein annotations. Then, we applied the group-based phosphorylation scoring

(GPS) method on the data set; here, we present a comprehensive PK-specific

prediction server GPS, which could predict kinase-specific phosphorylation sites

from protein primary sequences for 71 different PK groups. GPS has been

implemented in PHP and is available on a www server at

http://973-proteinweb.ustc.edu.cn/gps/gps\_web/.

DOI: 10.1093/nar/gki393

PMCID: PMC1160154

PMID: 15980451 [Indexed for MEDLINE]

3143. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W180-3.

PREDBALB/c: a system for the prediction of peptide binding to H2d molecules, a

haplotype of the BALB/c mouse.

Zhang GL(1), Srinivasan KN, Veeramani A, August JT, Brusic V.

Author information:

(1)Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613.

PRED(BALB/c) is a computational system that predicts peptides binding to the

major histocompatibility complex-2 (H2(d)) of the BALB/c mouse, an important

laboratory model organism. The predictions include the complete set of H2(d)

class I (H2-K(d), H2-L(d) and H2-D(d)) and class II (I-E(d) and I-A(d))

molecules. The prediction system utilizes quantitative matrices, which were

rigorously validated using experimentally determined binders and non-binders and

also by in vivo studies using viral proteins. The prediction performance of

PRED(BALB/c) is of very high accuracy. To our knowledge, this is the first online

server for the prediction of peptides binding to a complete set of major

histocompatibility complex molecules in a model organism (H2(d) haplotype).

PRED(BALB/c) is available at http://antigen.i2r.a-star.edu.sg/predBalbc/.

DOI: 10.1093/nar/gki479

PMCID: PMC1160239

PMID: 15980450 [Indexed for MEDLINE]

3144. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W168-71.

CEP: a conformational epitope prediction server.

Kulkarni-Kale U(1), Bhosle S, Kolaskar AS.

Author information:

(1)Bioinformatics Centre, University of Pune, Pune 411 007, India.

CEP server (http://bioinfo.ernet.in/cep.htm) provides a web interface to the

conformational epitope prediction algorithm developed in-house. The algorithm,

apart from predicting conformational epitopes, also predicts antigenic

determinants and sequential epitopes. The epitopes are predicted using 3D

structure data of protein antigens, which can be visualized graphically. The

algorithm employs structure-based Bioinformatics approach and solvent

accessibility of amino acids in an explicit manner. Accuracy of the algorithm was

found to be 75% when evaluated using X-ray crystal structures of Ag-Ab complexes

available in the PDB. This is the first and the only method available for the

prediction of conformational epitopes, which is an attempt to map probable

antibody-binding sites of protein antigens.

DOI: 10.1093/nar/gki460

PMCID: PMC1160221

PMID: 15980448 [Indexed for MEDLINE]

3145. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W164-7.

TMBETA-NET: discrimination and prediction of membrane spanning beta-strands in

outer membrane proteins.

Gromiha MM(1), Ahmad S, Suwa M.

Author information:

(1)Computational Biology Research Center, National Institute of Advanced

Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research

Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan. michael-gromiha@aist.go.jp

We have developed a web-server, TMBETA-NET for discriminating outer membrane

proteins and predicting their membrane spanning beta-strand segments. The amino

acid compositions of globular and outer membrane proteins have been

systematically analyzed and a statistical method has been proposed for

discriminating outer membrane proteins. The prediction of membrane spanning

segments is mainly based on feed forward neural network and refined with

beta-strand length. Our program takes the amino acid sequence as input and

displays the type of the protein along with membrane-spanning beta-strand

segments as a stretch of highlighted amino acid residues. Further, the

probability of residues to be in transmembrane beta-strand has been provided with

a coloring scheme. We observed that outer membrane proteins were discriminated

with an accuracy of 89% and their membrane spanning beta-strand segments at an

accuracy of 73% just from amino acid sequence information. The prediction server

is available at http://psfs.cbrc.jp/tmbeta-net/.

DOI: 10.1093/nar/gki367

PMCID: PMC1160128

PMID: 15980447 [Indexed for MEDLINE]

3146. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W143-7.

GPCRsclass: a web tool for the classification of amine type of G-protein-coupled

receptors.

Bhasin M(1), Raghava GP.

Author information:

(1)Bioinformatics Center, Institute of Microbial Technology, Sector 39A,

Chandigarh, India.

The receptors of amine subfamily are specifically major drug targets for therapy

of nervous disorders and psychiatric diseases. The recognition of novel amine

type of receptors and their cognate ligands is of paramount interest for

pharmaceutical companies. In the past, Chou and co-workers have shown that

different types of amine receptors are correlated with their amino acid

composition and are predictable on its basis with considerable accuracy [Elrod

and Chou (2002) Protein Eng., 15, 713-715]. This motivated us to develop a better

method for the recognition of novel amine receptors and for their further

classification. The method was developed on the basis of amino acid composition

and dipeptide composition of proteins using support vector machine. The method

was trained and tested on 167 proteins of amine subfamily of G-protein-coupled

receptors (GPCRs). The method discriminated amine subfamily of GPCRs from

globular proteins with Matthew's correlation coefficient of 0.98 and 0.99 using

amino acid composition and dipeptide composition, respectively. In classifying

different types of amine receptors using amino acid composition and dipeptide

composition, the method achieved an accuracy of 89.8 and 96.4%, respectively. The

performance of the method was evaluated using 5-fold cross-validation. The

dipeptide composition based method predicted 67.6% of protein sequences with an

accuracy of 100% with a reliability index > or =5. A web server GPCRsclass has

been developed for predicting amine-binding receptors from its amino acid

sequence [http://www.imtech.res.in/raghava/gpcrsclass/ and

http://bioinformatics.uams.edu/raghava/gpersclass/ (mirror site)].

DOI: 10.1093/nar/gki351

PMCID: PMC1160112

PMID: 15980444 [Indexed for MEDLINE]

3147. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W138-42.

PEPVAC: a web server for multi-epitope vaccine development based on the

prediction of supertypic MHC ligands.

Reche PA(1), Reinherz EL.

Author information:

(1)Laboratory of Immunobiology and Department of Medical Oncology, Dana-Farber

Cancer Institute, Harvard Medical School, 44 Binney Street, Boston, MA 02115,

USA. reche@research.dfci.harvard.edu

Prediction of peptide binding to major histocompatibility complex (MHC) molecules

is a basis for anticipating T-cell epitopes, as well as epitope discovery-driven

vaccine development. In the human, MHC molecules are known as human leukocyte

antigens (HLAs) and are extremely polymorphic. HLA polymorphism is the basis of

differential peptide binding, until now limiting the practical use of current

epitope-prediction tools for vaccine development. Here, we describe a web server,

PEPVAC (Promiscuous EPitope-based VACcine), optimized for the formulation of

multi-epitope vaccines with broad population coverage. This optimization is

accomplished through the prediction of peptides that bind to several HLA

molecules with similar peptide-binding specificity (supertypes). Specifically, we

offer the possibility of identifying promiscuous peptide binders to five distinct

HLA class I supertypes (A2, A3, B7, A24 and B15). We estimated the phenotypic

population frequency of these supertypes to be 95%, regardless of ethnicity.

Targeting these supertypes for promiscuous peptide-binding predictions results in

a limited number of potential epitopes without compromising the population

coverage required for practical vaccine design considerations. PEPVAC can also

identify conserved MHC ligands, as well as those with a C-terminus resulting from

proteasomal cleavage. The combination of these features with the prediction of

promiscuous HLA class I ligands further limits the number of potential epitopes.

The PEPVAC server is hosted by the Dana-Farber Cancer Institute at the site

http://immunax.dfci.harvard.edu/PEPVAC/.

DOI: 10.1093/nar/gki357

PMCID: PMC1160118

PMID: 15980443 [Indexed for MEDLINE]

3148. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W133-7.

pdbFun: mass selection and fast comparison of annotated PDB residues.

Ausiello G(1), Zanzoni A, Peluso D, Via A, Helmer-Citterich M.

Author information:

(1)Centre for Molecular Bioinformatics, Department of Biology, University of Rome

Tor Vergata Via della Ricerca Scientifica, 00133 Rome, Italy.

gabriele@cbm.bio.uniroma2.it

pdbFun (http://pdbfun.uniroma2.it) is a web server for structural and functional

analysis of proteins at the residue level. pdbFun gives fast access to the whole

Protein Data Bank (PDB) organized as a database of annotated residues. The

available data (features) range from solvent exposure to ligand binding ability,

location in a protein cavity, secondary structure, residue type, sequence

functional pattern, protein domain and catalytic activity. Users can select any

residue subset (even including any number of PDB structures) by combining the

available features. Selections can be used as probe and target in multiple

structure comparison searches. For example a search could involve, as a query,

all solvent-exposed, hydrophylic residues that are not in alpha-helices and are

involved in nucleotide binding. Possible examples of targets are represented by

another selection, a single structure or a dataset composed of many structures.

The output is a list of aligned structural matches offered in tabular and also

graphical format.

DOI: 10.1093/nar/gki499

PMCID: PMC1160259

PMID: 15980442 [Indexed for MEDLINE]

3149. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W130-2.

DIAL: a web-based server for the automatic identification of structural domains

in proteins.

Pugalenthi G(1), Archunan G, Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences, Tata Institute of Fundamental

Research UAS-GKVK campus, Bellary Road, Bangalore 560 065, Karnataka, India.

DIAL is a web server for the automatic identification of structural domains given

the 3D coordinates of a protein. Delineation of the structural domains and their

exact boundaries are the starting points for the better realization of distantly

related members of the domain families, for the rational design of the

experiments and for clearer understanding of the biological function. The current

server can examine crystallographic multiple chains and provide structural domain

solutions that can also describe domain swapping events. The server can be

accessed from http://www.ncbs.res.in/~faculty/mini/DIAL/home.html. The

Supplementary data can be accessed from

http://www.ncbs.res.in/~faculty/mini/DIAL/supplement.html.

DOI: 10.1093/nar/gki427

PMCID: PMC1160188

PMID: 15980441 [Indexed for MEDLINE]

3150. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W105-10.

LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using

SVM and profile of PSI-BLAST.

Xie D(1), Li A, Wang M, Fan Z, Feng H.

Author information:

(1)Department of Electronic Science and Technology, University of Science and

Technology of China, Hefei, People's Republic of China.

Subcellular location of a protein is one of the key functional characters as

proteins must be localized correctly at the subcellular level to have normal

biological function. In this paper, a novel method named LOCSVMPSI has been

introduced, which is based on the support vector machine (SVM) and the

position-specific scoring matrix generated from profiles of PSI-BLAST. With a

jackknife test on the RH2427 data set, LOCSVMPSI achieved a high overall

prediction accuracy of 90.2%, which is higher than the prediction results by

SubLoc and ESLpred on this data set. In addition, prediction performance of

LOCSVMPSI was evaluated with 5-fold cross validation test on the PK7579 data set

and the prediction results were consistently better than the previous method

based on several SVMs using composition of both amino acids and amino acid pairs.

Further test on the SWISSPROT new-unique data set showed that LOCSVMPSI also

performed better than some widely used prediction methods, such as PSORTII,

TargetP and LOCnet. All these results indicate that LOCSVMPSI is a powerful tool

for the prediction of eukaryotic protein subcellular localization. An online web

server (current version is 1.3) based on this method has been developed and is

freely available to both academic and commercial users, which can be accessed by

at http://Bioinformatics.ustc.edu.cn/LOCSVMPSI/LOCSVMPSI.php.

DOI: 10.1093/nar/gki359

PMCID: PMC1160120

PMID: 15980436 [Indexed for MEDLINE]

3151. BMC Bioinformatics. 2005 Jun 29;6:164.

Species-specific analysis of protein sequence motifs using mutual information.

Hummel J(1), Keshvari N, Weckwerth W, Selbig J.

Author information:

(1)Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14424

Potsdam, Germany. hummel@mpimp-golm.mpg.de

BACKGROUND: Protein sequence motifs are by definition short fragments of

conserved amino acids, often associated with a specific function. Accordingly

protein sequence profiles derived from multiple sequence alignments provide an

alternative description of functional motifs characterizing families of related

sequences. Such profiles conveniently reflect functional necessities by pointing

out proximity at conserved sequence positions as well as depicting distances at

variable positions. Discovering significant conservation characteristics within

the variable positions of profiles mirrors group-specific and, in particular,

evolutionary features of the underlying sequences.

RESULTS: We describe the tool PROfile analysis based on Mutual Information

(PROMI) that enables comparative analysis of user-classified protein sequences.

PROMI is implemented as a web service using Perl and R as well as other publicly

available packages and tools on the server-side. On the client-side

platform-independence is achieved by generally applied internet delivery

standards. As one possible application analysis of the zinc finger C2H2-type

protein domain is introduced to illustrate the functionality of the tool.

CONCLUSION: The web service PROMI should assist researchers to detect

evolutionary correlations in protein profiles of defined biological sequences. It

is available at http://promi.mpimp-golm.mpg.de where additional documentation can

be found.

DOI: 10.1186/1471-2105-6-164

PMCID: PMC1182352

PMID: 15987530 [Indexed for MEDLINE]

3152. BMC Bioinformatics. 2005 Jun 17;6:152.

pSLIP: SVM based protein subcellular localization prediction using multiple

physicochemical properties.

Sarda D(1), Chua GH, Li KB, Krishnan A.

Author information:

(1)Bioinformatics Institute, 30, Biopolis Street, #07-01, 138671, Singapore.

deepak@bii.a-star.edu.sg

BACKGROUND: Protein subcellular localization is an important determinant of

protein function and hence, reliable methods for prediction of localization are

needed. A number of prediction algorithms have been developed based on amino acid

compositions or on the N-terminal characteristics (signal peptides) of proteins.

However, such approaches lead to a loss of contextual information. Moreover,

where information about the physicochemical properties of amino acids has been

used, the methods employed to exploit that information are less than optimal and

could use the information more effectively.

RESULTS: In this paper, we propose a new algorithm called pSLIP which uses

Support Vector Machines (SVMs) in conjunction with multiple physicochemical

properties of amino acids to predict protein subcellular localization in

eukaryotes across six different locations, namely, chloroplast, cytoplasmic,

extracellular, mitochondrial, nuclear and plasma membrane. The algorithm was

applied to the dataset provided by Park and Kanehisa and we obtained prediction

accuracies for the different classes ranging from 87.7%-97.0% with an overall

accuracy of 93.1%.

CONCLUSION: This study presents a physicochemical property based protein

localization prediction algorithm. Unlike other algorithms, contextual

information is preserved by dividing the protein sequences into clusters. The

prediction accuracy shows an improvement over other algorithms based on various

types of amino acid composition (single, pair and gapped pair). We have also

implemented a web server to predict protein localization across the six classes

(available at http://pslip.bii.a-star.edu.sg/).

DOI: 10.1186/1471-2105-6-152

PMCID: PMC1182350

PMID: 15963230 [Indexed for MEDLINE]

3153. Bioinformatics. 2005 Jun 15;21(12):2917-20. Epub 2005 Apr 19.

Meta-DP: domain prediction meta-server.

Saini HK(1), Fischer D.

Author information:

(1)Center of Excellence in Bioinformatics and Department of Computer Science and

Engineering, University at Buffalo, 901 Washington Street, Suite 300, Buffalo, NY

14203, USA. hkaur@bioinformatics.buffalo.edu

Meta-DP, a domain prediction meta-server provides a simple interface to predict

domains in a given protein sequence using a number of domain prediction methods.

The Meta-DP is a convenient resource because through accessing a single site,

users automatically obtain the results of the various domain prediction methods

along with a consensus prediction. The Meta-DP is currently coupled to 10 domain

prediction servers and can be extended to include any number of methods. Meta-DP

can thus become a centralized repository of available methods. Meta-DP was also

used to evaluate the performance of 13 domain prediction methods in the context

of CAFASP-DP.AVAILABILITY: The Meta-DP server is freely available at

http://meta-dp.bioinformatics.buffalo.edu and the CAFASP-DP evaluation results

are available at http://cafasp4.bioinformatics.buffalo.edu/dp/update.html

CONTACT: hkaur@bioinformatics.buffalo.edu

SUPPLEMENTARY INFORMATION: Available at

http://cafasp4.bioinformatics.buffalo.edu/dp/update.html.

DOI: 10.1093/bioinformatics/bti445

PMID: 15840708 [Indexed for MEDLINE]

3154. Bioinformatics. 2005 Jun;21 Suppl 1:i251-7.

An HMM posterior decoder for sequence feature prediction that includes homology

information.

Käll L(1), Krogh A, Sonnhammer EL.

Author information:

(1)Center for Genomics and Bioinformatics, Karolinska Institutet SE-17 177

Stockholm, Sweden.

MOTIVATION: When predicting sequence features like transmembrane topology, signal

peptides, coil-coil structures, protein secondary structure or genes, extra

support can be gained from homologs.

RESULTS: We present here a general hidden Markov model (HMM) decoding algorithm

that combines probabilities for sequence features of homologs by considering the

average of the posterior label probability of each position in a global sequence

alignment. The algorithm is an extension of the previously described 'optimal

accuracy' decoder, allowing homology information to be used. It was benchmarked

using an HMM for transmembrane topology and signal peptide prediction, Phobius.

We found that the performance was substantially increased when incorporating

information from homologs.

AVAILABILITY: A prediction server for transmembrane topology and signal peptides

that uses the algorithm is available at http://phobius.cgb.ki.se/poly.html. An

implementation of the algorithm is available on request from the authors.

DOI: 10.1093/bioinformatics/bti1014

PMID: 15961464 [Indexed for MEDLINE]

3155. Immunogenetics. 2005 Jun;57(5):304-14. Epub 2005 May 3.

Automated generation and evaluation of specific MHC binding predictive tools: ARB

matrix applications.

Bui HH(1), Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton KA, Mothé BR,

Chisari FV, Watkins DI, Sette A.

Author information:

(1)Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology,

3030 Bunker Hill Street, Suite 326, San Diego, CA 92109, USA.

Prediction of which peptides can bind major histocompatibility complex (MHC)

molecules is commonly used to assist in the identification of T cell epitopes.

However, because of the large numbers of different MHC molecules of interest,

each associated with different predictive tools, tool generation and evaluation

can be a very resource intensive task. A methodology commonly used to predict MHC

binding affinity is the matrix or linear coefficients method. Herein, we

described Average Relative Binding (ARB) matrix methods that directly predict

IC(50) values allowing combination of searches involving different peptide sizes

and alleles into a single global prediction. A computer program was developed to

automate the generation and evaluation of ARB predictive tools. Using an in-house

MHC binding database, we generated a total of 85 and 13 MHC class I and class II

matrices, respectively. Results from the automated evaluation of tool efficiency

are presented. We anticipate that this automation framework will be generally

applicable to the generation and evaluation of large numbers of MHC predictive

methods and tools, and will be of value to centralize and rationalize the process

of evaluation of MHC predictions. MHC binding predictions based on ARB matrices

were made available at http://epitope.liai.org:8080/matrix web server.

DOI: 10.1007/s00251-005-0798-y

PMID: 15868141 [Indexed for MEDLINE]

3156. J Comput Aided Mol Des. 2005 Jun;19(6):453-63.

Virtual computational chemistry laboratory--design and description.

Tetko IV(1), Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin

VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV.

Author information:

(1)Institute of Bioorganic & Petroleum Chemistry, Kyiv, Ukraine.

itetko@vcclab.org

Internet technology offers an excellent opportunity for the development of tools

by the cooperative effort of various groups and institutions. We have developed a

multi-platform software system, Virtual Computational Chemistry Laboratory,

http://www.vcclab.org, allowing the computational chemist to perform a

comprehensive series of molecular indices/properties calculations and data

analysis. The implemented software is based on a three-tier architecture that is

one of the standard technologies to provide client-server services on the

Internet. The developed software includes several popular programs, including the

indices generation program, DRAGON, a 3D structure generator, CORINA, a program

to predict lipophilicity and aqueous solubility of chemicals, ALOGPS and others.

All these programs are running at the host institutes located in five countries

over Europe. In this article we review the main features and statistics of the

developed system that can be used as a prototype for academic and industry

models.

DOI: 10.1007/s10822-005-8694-y

PMID: 16231203 [Indexed for MEDLINE]

3157. Proteins. 2005 Jun 1;59(4):828-39.

High accuracy prediction of beta-turns and their types using propensities and

multiple alignments.

Fuchs PF(1), Alix AJ.

Author information:

(1)Equipe de Bioinformatique Génomique et Moléculaire, INSERM EMI 0346,

Université Paris 7, Paris, France. Patrick.Fuchs@ebgm.jussieu.fr

We have developed a method that predicts both the presence and the type of

beta-turns, using a straightforward approach based on propensities and multiple

alignments. The propensities were calculated classically, but the way to use them

for prediction was completely new: starting from a tetrapeptide sequence on which

one wants to evaluate the presence of a beta-turn, the propensity for a given

residue is modified by taking into account all the residues present in the

multiple alignment at this position. The evaluation of a score is then done by

weighting these propensities by the use of Position-specific score matrices

generated by PSI-BLAST. The introduction of secondary structure information

predicted by PSIPRED or SSPRO2 as well as taking into account the flanking

residues around the tetrapeptide improved the accuracy greatly. This latter

evaluated on a database of 426 reference proteins (previously used on other

studies) by a sevenfold crossvalidation gave very good results with a Matthews

Correlation Coefficient (MCC) of 0.42 and an overall prediction accuracy of

74.8%; this places our method among the best ones. A jackknife test was also

done, which gave results within the same range. This shows that it is possible to

reach neural networks accuracy with considerably less computional cost and

complexity. Furthermore, propensities remain excellent descriptors of amino acid

tendencies to belong to beta-turns, which can be useful for peptide or protein

engineering and design. For beta-turn type prediction, we reached the best

accuracy ever published in terms of MCC (except for the irregular type IV) in the

range of 0.25-0.30 for types I, II, and I' and 0.13-0.15 for types VIII, II', and

IV. To our knowledge, our method is the only one available on the Web that

predicts types I' and II'. The accuracy evaluated on two larger databases of 547

and 823 proteins was not improved significantly. All of this was implemented into

a Web server called COUDES (French acronym for: Chercher Ou Une Deviation Existe

Surement), which is available at the following URL:

http://bioserv.rpbs.jussieu.fr/Coudes/index.html within the new bioinformatics

platform RPBS.

DOI: 10.1002/prot.20461

PMID: 15822097 [Indexed for MEDLINE]

3158. BMC Bioinformatics. 2005 May 23;6:124.

arrayCGHbase: an analysis platform for comparative genomic hybridization

microarrays.

Menten B(1), Pattyn F, De Preter K, Robbrecht P, Michels E, Buysse K, Mortier G,

De Paepe A, van Vooren S, Vermeesch J, Moreau Y, De Moor B, Vermeulen S, Speleman

F, Vandesompele J.

Author information:

(1)Center for Medical Genetics, Ghent University Hospital, De Pintelaan 185,

B-9000 Ghent, Belgium. bjorn.menten@ugent.be <bjorn.menten@ugent.be>

BACKGROUND: The availability of the human genome sequence as well as the large

number of physically accessible oligonucleotides, cDNA, and BAC clones across the

entire genome has triggered and accelerated the use of several platforms for

analysis of DNA copy number changes, amongst others microarray comparative

genomic hybridization (arrayCGH). One of the challenges inherent to this new

technology is the management and analysis of large numbers of data points

generated in each individual experiment.

RESULTS: We have developed arrayCGHbase, a comprehensive analysis platform for

arrayCGH experiments consisting of a MIAME (Minimal Information About a

Microarray Experiment) supportive database using MySQL underlying a data mining

web tool, to store, analyze, interpret, compare, and visualize arrayCGH results

in a uniform and user-friendly format. Following its flexible design,

arrayCGHbase is compatible with all existing and forthcoming arrayCGH platforms.

Data can be exported in a multitude of formats, including BED files to map copy

number information on the genome using the Ensembl or UCSC genome browser.

CONCLUSION: ArrayCGHbase is a web based and platform independent arrayCGH data

analysis tool, that allows users to access the analysis suite through the

internet or a local intranet after installation on a private server. ArrayCGHbase

is available at http://medgen.ugent.be/arrayCGHbase/.

DOI: 10.1186/1471-2105-6-124

PMCID: PMC1173083

PMID: 15910681 [Indexed for MEDLINE]

3159. Bioinformatics. 2005 May 15;21(10):2287-93. Epub 2005 Mar 29.

Quasi-consensus-based comparison of profile hidden Markov models for protein

sequences.

Kahsay RY(1), Wang G, Gao G, Liao L, Dunbrack R.

Author information:

(1)Delaware Biotechnology Institute, Newark, DE 19715, USA.

A simple approach for the sensitive detection of distant relationships among

protein families and for sequence-structure alignment via comparison of hidden

Markov models based on their quasi-consensus sequences is presented. Using a

previously published benchmark dataset, the approach is demonstrated to give

better homology detection and yield alignments with improved accuracy in

comparison to an existing state-of-the-art dynamic programming profile-profile

comparison method. This method also runs significantly faster and is therefore

suitable for a server covering the rapidly increasing structure database. A

server based on this method is available at

http://liao.cis.udel.edu/website/servers/modmod

DOI: 10.1093/bioinformatics/bti374

PMID: 15797916 [Indexed for MEDLINE]

3160. Bioinformatics. 2005 May 15;21(10):2322-8. Epub 2005 Mar 15.

Identification and measurement of neighbor-dependent nucleotide substitution

processes.

Arndt PF(1), Hwa T.

Author information:

(1)Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin,

Germany. arndt@molgen.mpg.de

MOTIVATION: Neighbor-dependent substitution processes generated specific pattern

of dinucleotide frequencies in the genomes of most organisms. The

CpG-methylation-deamination process is, e.g. a prominent process in vertebrates

(CpG effect). Such processes, often with unknown mechanistic origins, need to be

incorporated into realistic models of nucleotide substitutions.

RESULTS: Based on a general framework of nucleotide substitutions we developed a

method that is able to identify the most relevant neighbor-dependent substitution

processes, estimate their relative frequencies and judge their importance in

order to be included into the modeling. Starting from a model for neighbor

independent nucleotide substitution we successively added neighbor-dependent

substitution processes in the order of their ability to increase the likelihood

of the model describing given data. The analysis of neighbor-dependent nucleotide

substitutions based on repetitive elements found in the genomes of human,

zebrafish and fruit fly is presented.

AVAILABILITY: A web server to perform the presented analysis is freely available

at: http://evogen.molgen.mpg.de/server/substitution-analysis

DOI: 10.1093/bioinformatics/bti376

PMID: 15769841 [Indexed for MEDLINE]

3161. Bioinformatics. 2005 May 15;21(10):2541-3. Epub 2005 Mar 3.

PSIbase: a database of Protein Structural Interactome map (PSIMAP).

Gong S(1), Yoon G, Jang I, Bolser D, Dafas P, Schroeder M, Choi H, Cho Y, Han K,

Lee S, Choi H, Lappe M, Holm L, Kim S, Oh D, Bhak J.

Author information:

(1)Biomatics Lab, Department of BioSystems, KAIST, Daejeon, Korea.

Protein Structural Interactome map (PSIMAP) is a global interaction map that

describes domain-domain and protein-protein interaction information for known

Protein Data Bank structures. It calculates the Euclidean distance to determine

interactions between possible pairs of structural domains in proteins. PSIbase is

a database and file server for protein structural interaction information

calculated by the PSIMAP algorithm. PSIbase also provides an easy-to-use protein

domain assignment module, interaction navigation and visual tools. Users can

retrieve possible interaction partners of their proteins of interests if a

significant homology assignment is made with their query sequences.AVAILABILITY:

http://psimap.org and http://psibase.kaist.ac.kr/

DOI: 10.1093/bioinformatics/bti366

PMID: 15749693 [Indexed for MEDLINE]

3162. Bioinformatics. 2005 May 15;21(10):2362-9. Epub 2005 Mar 3.

Multiple flexible structure alignment using partial order graphs.

Ye Y(1), Godzik A.

Author information:

(1)Program in Bioinformatics and Systems Biology, The Burnham Institute, 10901 N.

Torrey Pines Road, La Jolla, CA 92037, USA. yye@burnham.org

MOTIVATION: Existing comparisons of protein structures are not able to describe

structural divergence and flexibility in the structures being compared because

they focus on identifying a common invariant core and ignore parts of the

structures outside this core. Understanding the structural divergence and

flexibility is critical for studying the evolution of functions and specificities

of proteins.

RESULTS: A new method of multiple protein structure alignment, POSA (Partial

Order Structure Alignment), was developed using a partial order graph

representation of multiple alignments. POSA has two unique features: (1)

identifies and classifies regions that are conserved only in a subset of input

structures and (2) allows internal rearrangements in protein structures. POSA

outperforms other programs in the cases where structural flexibilities exist and

provides new insights by visualizing the mosaic nature of multiple structural

alignments. POSA is an ideal tool for studying the variation of protein

structures within diverse structural families.

AVAILABILITY: POSA is freely available for academic users on a Web server at

http://fatcat.burnham.org/POSA

DOI: 10.1093/bioinformatics/bti353

PMID: 15746292 [Indexed for MEDLINE]

3163. Bioinformatics. 2005 May 15;21(10):2550-1. Epub 2005 Mar 3.

Gene-Expression Omnibus integration and clustering tools in SeqExpress.

Boyle J(1).

Author information:

(1)john@seqexpress.com

SeqExpress, a gene-expression analysis suite, has been extended to offer a number

of cluster generation, refinement and visualization techniques. The cluster

generation methods have been specialized to deal with aspects of the sparseness

and extreme values that occur within microarray data. The results of such cluster

analysis can then be refined using either: a functional enrichment based

procedure, which examines each cluster to see if it possesses an unusually high

or low concentration of ontology terms; or by using Expectation-Maximization to

find a mixture of model based distributions within the datasets. Visualizations

are provided both to explore and compare the results of the cluster generation

algorithms. In addition, a tool has been developed which integrates SeqExpress

with the Gene-Expression Omnibus repository. The tool provides seamless access to

the large number of experimental results in the repository, so that they can be

visualized and analysed locally using SeqExpress.AVAILABILITY: SeqExpress is

available as a 6 MB download from http://www.seqexpress.com and runs under

Windows. A server-based version is available and is required for the GEO

integration. SeqExpress is not affiliated with any academic institution, funding

body or commercial organization and is free to use by all.

DOI: 10.1093/bioinformatics/bti355

PMID: 15746290 [Indexed for MEDLINE]

3164. Bioinformatics. 2005 May 15;21(10):2570-1. Epub 2005 Mar 3.

WebAllergen: a web server for predicting allergenic proteins.

Riaz T(1), Hor HL, Krishnan A, Tang F, Li KB.

Author information:

(1)Bioinformatics Institute, 30 Biopolis Street, 138671, Singapore.

WebAllergen is a web server that predicts the potential allergenicity of

proteins. The query protein will be compared against a set of prebuilt allergenic

motifs that have been obtained from 664 known allergen proteins. The query will

also be compared with known allergens that do not have detectable allergenic

motifs. Moreover, users are allowed to upload their own allergens as alternative

training sequences on which a new set of allergenic motifs will be built. The

query sequences can also be compared with these motifs.AVAILABILITY:

http://weballergen.bii.a-star.edu.sg/

DOI: 10.1093/bioinformatics/bti356

PMID: 15746289 [Indexed for MEDLINE]

3165. Bioinformatics. 2005 May 15;21(10):2525-7. Epub 2005 Feb 22.

AutoMotif server: prediction of single residue post-translational modifications

in proteins.

Plewczynski D(1), Tkacz A, Wyrwicz LS, Rychlewski L.

Author information:

(1)BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland.

darman@bioinfo.pl

The AutoMotif Server allows for identification of post-translational modification

(PTM) sites in proteins based only on local sequence information. The local

sequence preferences of short segments around PTM residues are described here as

linear functional motifs (LFMs). Sequence models for all types of PTMs are

trained by support vector machine on short-sequence fragments of proteins in the

current release of Swiss-Prot database (phosphorylation by various protein

kinases, sulfation, acetylation, methylation, amidation, etc.). The accuracy of

the identification is estimated using the standard leave-one-out procedure. The

sensitivities for all types of short LFMs are in the range of 70%.AVAILABILITY:

The AutoMotif Server is available free for academic use at

http://automotif.bioinfo.pl/

DOI: 10.1093/bioinformatics/bti333

PMID: 15728119 [Indexed for MEDLINE]

3166. Bioinformatics. 2005 May 15;21(10):2522-4. Epub 2005 Feb 4.

PSLpred: prediction of subcellular localization of bacterial proteins.

Bhasin M(1), Garg A, Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector 39A, Chandigarh, India.

SUMMARY: We developed a web server PSLpred for predicting subcellular

localization of gram-negative bacterial proteins with an overall accuracy of

91.2%. PSLpred is a hybrid approach-based method that integrates PSI-BLAST and

three SVM modules based on compositions of residues, dipeptides and

physico-chemical properties. The prediction accuracies of 90.7, 86.8, 90.3, 95.2

and 90.6% were attained for cytoplasmic, extracellular, inner-membrane,

outer-membrane and periplasmic proteins, respectively. Furthermore, PSLpred was

able to predict approximately 74% of sequences with an average prediction

accuracy of 98% at RI = 5.

AVAILABILITY: PSLpred is available at http://www.imtech.res.in/raghava/pslpred/

DOI: 10.1093/bioinformatics/bti309

PMID: 15699023 [Indexed for MEDLINE]

3167. Bioinformation. 2005 May 15;1(1):21-4.

T-Epitope Designer: A HLA-peptide binding prediction server.

Kangueane P(1), Sakharkar MK.

Author information:

(1)School of Mechanical and Aerospace Engineering, Nanyang Technological

University, Singapore 639798. mpandjassarame@ntu.edu.sg

The current challenge in synthetic vaccine design is the development of a

methodology to identify and test short antigen peptides as potential T-cell

epitopes. Recently, we described a HLA-peptide binding model (using structural

properties) capable of predicting peptides binding to any HLA allele.

Consequently, we have developed a web server named T-EPITOPE DESIGNER to

facilitate HLA-peptide binding prediction. The prediction server is based on a

model that defines peptide binding pockets using information gleaned from X-ray

crystal structures of HLA-peptide complexes, followed by the estimation of

peptide binding to binding pockets. Thus, the prediction server enables the

calculation of peptide binding to HLA alleles. This model is superior to many

existing methods because of its potential application to any given HLA allele

whose sequence is clearly defined. The web server finds potential application in

T cell epitope vaccine design.AVAILABILITY: http://www.bioinformation.net/ted/

PMCID: PMC1891623

PMID: 17597847

3168. Proteins. 2005 May 15;59(3):467-75.

Combining prediction of secondary structure and solvent accessibility in

proteins.

Adamczak R(1), Porollo A, Meller J.

Author information:

(1)Biomedical Informatics, Children's Hospital Research Foundation, Cincinnati,

Ohio 45229, USA.

Owing to the use of evolutionary information and advanced machine learning

protocols, secondary structures of amino acid residues in proteins can be

predicted from the primary sequence with more than 75% per-residue accuracy for

the 3-state (i.e., helix, beta-strand, and coil) classification problem. In this

work we investigate whether further progress may be achieved by incorporating the

relative solvent accessibility (RSA) of an amino acid residue as a fingerprint of

the overall topology of the protein. Toward that goal, we developed a novel

method for secondary structure prediction that uses predicted RSA in addition to

attributes derived from evolutionary profiles. Our general approach follows the

2-stage protocol of Rost and Sander, with a number of Elman-type recurrent neural

networks (NNs) combined into a consensus predictor. The RSA is predicted using

our recently developed regression-based method that provides real-valued RSA,

with the overall correlation coefficients between the actual and predicted RSA of

about 0.66 in rigorous tests on independent control sets. Using the predicted

RSA, we were able to improve the performance of our secondary structure

prediction by up to 1.4% and achieved the overall per-residue accuracy between

77.0% and 78.4% for the 3-state classification problem on different control sets

comprising, together, 603 proteins without homology to proteins included in the

training. The effects of including solvent accessibility depend on the quality of

RSA prediction. In the limit of perfect prediction (i.e., when using the actual

RSA values derived from known protein structures), the accuracy of secondary

structure prediction increases by up to 4%. We also observed that projecting

real-valued RSA into 2 discrete classes with the commonly used threshold of 25%

RSA decreases the classification accuracy for secondary structure prediction.

While the level of improvement of secondary structure prediction may be different

for prediction protocols that implicitly account for RSA in other ways, we

conclude that an increase in the 3-state classification accuracy may be achieved

when combining RSA with a state-of-the-art protocol utilizing evolutionary

profiles. The new method is available through a Web server at

http://sable.cchmc.org.

Copyright 2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20441

PMID: 15768403 [Indexed for MEDLINE]

3169. BMC Bioinformatics. 2005 May 13;6:116.

Predicting functional sites with an automated algorithm suitable for

heterogeneous datasets.

La D(1), Livesay DR.

Author information:

(1)Department of Biological Sciences, California State Polytechnic University,

Pomona, California 91768, USA. dla@csupomona.edu

BACKGROUND: In a previous report (La et al., Proteins, 2005), we have

demonstrated that the identification of phylogenetic motifs, protein sequence

fragments conserving the overall familial phylogeny, represent a promising

approach for sequence/function annotation. Across a structurally and functionally

heterogeneous dataset, phylogenetic motifs have been demonstrated to correspond

to a wide variety of functional site archetypes, including those defined by

surface loops, active site clefts, and less exposed regions. However, in our

original demonstration of the technique, phylogenetic motif identification is

dependent upon a manually determined similarity threshold, prohibiting

large-scale application of the technique.

RESULTS: In this report, we present an algorithmic approach that determines

thresholds without human subjectivity. The approach relies on significant raw

data preprocessing to improve signal detection. Subsequently, Partition Around

Medoids Clustering (PAMC) of the similarity scores assesses sequence fragments

where functional annotation remains in question. The accuracy of the approach is

confirmed through comparisons to our previous (manual) results and structural

analyses. Triosephosphate isomerase and arginyl-tRNA synthetase are discussed as

exemplar cases. A quantitative functional site prediction assessment algorithm

indicates that the phylogenetic motif predictions, which require sequence

information only, are nearly as good as those from evolutionary trace methods

that do incorporate structure.

CONCLUSION: The automated threshold detection algorithm has been incorporated

into MINER, our web-based phylogenetic motif identification server. MINER is

freely available on the web at http://www.pmap.csupomona.edu/MINER/.

Pre-calculated functional site predictions of the COG database and an

implementation of the threshold detection algorithm, in the R statistical

language, can also be accessed at the website.

DOI: 10.1186/1471-2105-6-116

PMCID: PMC1142304

PMID: 15890082 [Indexed for MEDLINE]

3170. BMC Bioinformatics. 2005 May 12;6:114.

CAGER: classification analysis of gene expression regulation using multiple

information sources.

Ruan J(1), Zhang W.

Author information:

(1)Department of Computer Science and Engineering, Washington University, St.

Louis, MO 63130, USA. jruan@cse.wustl.edu

BACKGROUND: Many classification approaches have been applied to analyzing

transcriptional regulation of gene expressions. These methods build models that

can explain a gene's expression level from the regulatory elements (features) on

its promoter sequence. Different types of features, such as experimentally

verified binding motifs, motifs discovered by computer programs, or transcription

factor binding data measured with Chromatin Immunoprecipitation (ChIP) assays,

have been used towards this goal. Each type of features has been shown successful

in modeling gene transcriptional regulation under certain conditions. However, no

comparison has been made to evaluate the relative merit of these features.

Furthermore, most publicly available classification tools were not designed

specifically for modeling transcriptional regulation, and do not allow the user

to combine different types of features.

RESULTS: In this study, we use a specific classification method, decision trees,

to model transcriptional regulation in yeast with features based on predefined

motifs, automatically identified motifs, ChlP-chip data, or their combinations.

We compare the accuracies and stability of these models, and analyze their

capabilities in identifying functionally related genes. Furthermore, we design

and implement a user-friendly web server called CAGER (Classification Analysis of

Gene Expression Regulation) that integrates several software components for

automated analysis of transcriptional regulation using decision trees. Finally,

we use CAGER to study the transcriptional regulation of Arabidopsis genes in

response to abscisic acid, and report some interesting new results.

CONCLUSION: Models built with ChlP-chip data suffer from low accuracies when the

condition under which gene expressions are measured is significantly different

from the condition under which the ChIP experiment is conducted. Models built

with automatically identified motifs can sometimes discover new features, but

their modeling accuracies may have been over-estimated in previous studies.

Furthermore, models built with automatically identified motifs are not stable

with respect to noises. A combination of ChlP-chip data and predefined motifs can

substantially improve modeling accuracies, and is effective in identifying true

regulons. The CAGER web server, which is freely available at

http://cic.cs.wustl.edu/CAGER/, allows the user to select combinations of

different feature types for building decision trees, and interact with the models

graphically. We believe that it will be a useful tool to facilitate the discovery

of gene transcriptional regulatory networks.

DOI: 10.1186/1471-2105-6-114

PMCID: PMC1174863

PMID: 15890068 [Indexed for MEDLINE]

3171. Acta Crystallogr D Biol Crystallogr. 2005 May;61(Pt 5):634-6. Epub 2005 Apr 20.

SSEP-2.0: Secondary Structural Elements of Proteins.

Balamurugan B(1), Samaya Mohan K, Ramesh J, Roshan MN, Sumathi K, Sekar K.

Author information:

(1)Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India.

The Secondary Structural Elements of Proteins (SSEP) database is an integrated

and comprehensive knowledge base for accessing information related to all the

secondary-structural elements present in non-redundant (25 and 90%) protein

chains. The new version 2.0 of the database contains 2485 and 8595 protein chains

from the 25 and 90% non-redundant data sets, respectively. The necessary web

interfaces have been developed that enable users to visualize the

three-dimensional structure of the secondary-structural element in the client

machine using the free molecular-visualization program RASMOL. This source is

updated at regular intervals and can be accessed through the bioinformatics web

server at the URL http://cluster.physics.iisc.ernet.in/ssep or

http://144.16.71.148/ssep/.

DOI: 10.1107/S0907444905005883

PMID: 15858275 [Indexed for MEDLINE]

3172. Bioinformatics. 2005 May 1;21(9):2140-1. Epub 2005 Jan 18.

EPIMHC: a curated database of MHC-binding peptides for customized computational

vaccinology.

Reche PA(1), Zhang H, Glutting JP, Reinherz EL.

Author information:

(1)Laboratory of Immunobiology and Department of Medical Oncology, Dana-Farber

Cancer Institute, Boston, MA 02115, USA. reche@research.dfci.harvard.edu

SUMMARY: EPIMHC is a relational database of MHC-binding peptides and T cell

epitopes that are observed in real proteins. Currently, the database contains

4867 distinct peptide sequences from various sources, including 84

tumor-associated antigens. The EPIMHC database is accessible through a web server

that has been designed to facilitate research in computational vaccinology.

Importantly, peptides resulting from a query can be selected to derive specific

motif-matrices. Subsequently, these motif-matrices can be used in combination

with a dynamic algorithm for predicting MHC-binding peptides from user-provided

protein queries.

AVAILABILITY: The EPIMHC database server is hosted by the Dana-Farber Cancer

Institute at the site http://immunax.dfci.harvard.edu/bioinformatics/epimhc/

DOI: 10.1093/bioinformatics/bti269

PMID: 15657103 [Indexed for MEDLINE]

3173. Bioinformatics. 2005 May 1;21(9):1901-7. Epub 2005 Jan 18.

PIBASE: a comprehensive database of structurally defined protein interfaces.

Davis FP(1), Sali A.

Author information:

(1)Graduate Group in Biophysics, California Institute for Quantitative Biomedical

Research, University of California, San Francisco, 94143, USA.

MOTIVATION: In recent years, the Protein Data Bank (PDB) has experienced rapid

growth. To maximize the utility of the high resolution protein-protein

interaction data stored in the PDB, we have developed PIBASE, a comprehensive

relational database of structurally defined interfaces between pairs of protein

domains. It is composed of binary interfaces extracted from structures in the PDB

and the Probable Quaternary Structure server using domain assignments from the

Structural Classification of Proteins and CATH fold classification systems.

RESULTS: PIBASE currently contains 158,915 interacting domain pairs between

105,061 domains from 2125 SCOP families. A diverse set of geometric,

physiochemical and topologic properties are calculated for each complex, its

domains, interfaces and binding sites. A subset of the interface properties are

used to remove interface redundancy within PDB entries, resulting in 20,912

distinct domain-domain interfaces. The complexes are grouped into 989 topological

classes based on their patterns of domain-domain contacts. The binary interfaces

and their corresponding binding sites are categorized into 18,755 and 30,975

topological classes, respectively, based on the topology of secondary structure

elements. The utility of the database is illustrated by outlining several current

applications.

AVAILABILITY: The database is accessible via the world wide web at

http://salilab.org/pibase

SUPPLEMENTARY INFORMATION: http://salilab.org/pibase/suppinfo.html.

DOI: 10.1093/bioinformatics/bti277

PMID: 15657096 [Indexed for MEDLINE]

3174. Bioinformatics. 2005 May 1;21(9):2101-3. Epub 2005 Jan 12.

Selecton: a server for detecting evolutionary forces at a single amino-acid site.

Doron-Faigenboim A(1), Stern A, Mayrose I, Bacharach E, Pupko T.

Author information:

(1)Department of Cell Research and Immunology, George S. Wise Faculty of Life

Sciences, Tel Aviv University, Ramat Aviv, Israel.

We present an algorithmic tool for the identification of biologically significant

amino acids in proteins of known three dimensional structure. We estimate the

degree of purifying selection and positive Darwinian selection at each site and

project these estimates onto the molecular surface of the protein. Thus, patches

of functional residues (undergoing either positive or purifying selection), which

may be discontinuous in the linear sequence, are revealed. We test for the

statistical significance of the site-specific scores in order to obtain reliable

and valid estimates.AVAILABILITY: The Selecton web server is available at:

http://selecton.bioinfo.tau.ac.il

SUPPLEMENTARY INFORMATION: More information is available at

http://selecton.bioinfo.tau.ac.il/overview.html. A set of examples is available

at http://selecton.bioinfo.tau.ac.il/gallery.html.

DOI: 10.1093/bioinformatics/bti259

PMID: 15647294 [Indexed for MEDLINE]

3175. BMC Bioinformatics. 2005 Apr 22;6:104.

A method for the prediction of GPCRs coupling specificity to G-proteins using

refined profile Hidden Markov Models.

Sgourakis NG(1), Bagos PG, Papasaikas PK, Hamodrakas SJ.

Author information:

(1)Department of Cell Biology and Biophysics, Faculty of Biology, University of

Athens, Athens 157 01, Greece. nsgourakis@biol.uoa.gr

BACKGROUND: G- Protein coupled receptors (GPCRs) comprise the largest group of

eukaryotic cell surface receptors with great pharmacological interest. A broad

range of native ligands interact and activate GPCRs, leading to signal

transduction within cells. Most of these responses are mediated through the

interaction of GPCRs with heterotrimeric GTP-binding proteins (G-proteins). Due

to the information explosion in biological sequence databases, the development of

software algorithms that could predict properties of GPCRs is important.

Experimental data reported in the literature suggest that heterotrimeric

G-proteins interact with parts of the activated receptor at the transmembrane

helix-intracellular loop interface. Utilizing this information and membrane

topology information, we have developed an intensive exploratory approach to

generate a refined library of statistical models (Hidden Markov Models) that

predict the coupling preference of GPCRs to heterotrimeric G-proteins. The method

predicts the coupling preferences of GPCRs to Gs, Gi/o and Gq/11, but not G12/13

subfamilies.

RESULTS: Using a dataset of 282 GPCR sequences of known coupling preference to

G-proteins and adopting a five-fold cross-validation procedure, the method

yielded an 89.7% correct classification rate. In a validation set comprised of

all receptor sequences that are species homologues to GPCRs with known coupling

preferences, excluding the sequences used to train the models, our method yields

a correct classification rate of 91.0%. Furthermore, promiscuous coupling

properties were correctly predicted for 6 of the 24 GPCRs that are known to

interact with more than one subfamily of G-proteins.

CONCLUSION: Our method demonstrates high correct classification rate. Unlike

previously published methods performing the same task, it does not require any

transmembrane topology prediction in a preceding step. A web-server for the

prediction of GPCRs coupling specificity to G-proteins available for

non-commercial users is located at http://bioinformatics.biol.uoa.gr/PRED-COUPLE.

DOI: 10.1186/1471-2105-6-104

PMCID: PMC1087828

PMID: 15847681 [Indexed for MEDLINE]

3176. BMC Bioinformatics. 2005 Apr 18;6:101.

MARS: microarray analysis, retrieval, and storage system.

Maurer M(1), Molidor R, Sturn A, Hartler J, Hackl H, Stocker G, Prokesch A,

Scheideler M, Trajanoski Z.

Author information:

(1)Institute for Genomics and Bioinformatics, Christian Doppler Laboratory for

Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010

Graz, Austria. michael.maurer@gmail.com

BACKGROUND: Microarray analysis has become a widely used technique for the study

of gene-expression patterns on a genomic scale. As more and more laboratories are

adopting microarray technology, there is a need for powerful and easy to use

microarray databases facilitating array fabrication, labeling, hybridization, and

data analysis. The wealth of data generated by this high throughput approach

renders adequate database and analysis tools crucial for the pursuit of insights

into the transcriptomic behavior of cells.

RESULTS: MARS (Microarray Analysis and Retrieval System) provides a comprehensive

MIAME supportive suite for storing, retrieving, and analyzing multi color

microarray data. The system comprises a laboratory information management system

(LIMS), a quality control management, as well as a sophisticated user management

system. MARS is fully integrated into an analytical pipeline of microarray image

analysis, normalization, gene expression clustering, and mapping of gene

expression data onto biological pathways. The incorporation of ontologies and the

use of MAGE-ML enables an export of studies stored in MARS to public repositories

and other databases accepting these documents.

CONCLUSION: We have developed an integrated system tailored to serve the specific

needs of microarray based research projects using a unique fusion of Web based

and standalone applications connected to the latest J2EE application server

technology. The presented system is freely available for academic and non-profit

institutions. More information can be found at http://genome.tugraz.at.

DOI: 10.1186/1471-2105-6-101

PMCID: PMC1090551

PMID: 15836795 [Indexed for MEDLINE]

3177. Bioinformatics. 2005 Apr 15;21(8):1727-9. Epub 2004 Dec 21.

Bloader--a batch loader application for MIAMExpress.

Schwager C(1), Blake J.

Author information:

(1)EMBL, Meyerhofstrasse 1, 67117 Heidelberg, Germany. schwager@embl.de

BLoader is a client server application for annotating and loading large amounts

of Microarray data into a local installation of the MIAMExpress database. A set

of nested spreadsheets is used to collect the required MIAME annotation.

Controlled vocabularies are downloaded from MIAMExpress and ArrayExpress

databases to guarantee MIAME compliance.AVAILABILITY: The application is

available from the author at http://www.ansorge-group.embl.de/bloader

CONTACT: schwager@embl.de

SUPPLEMENTARY INFORMATION: For more details on BLoader visit the above web page.

DOI: 10.1093/bioinformatics/bti231

PMID: 15613394 [Indexed for MEDLINE]

3178. Bioinformatics. 2005 Apr 15;21(8):1721-3. Epub 2004 Dec 21.

PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces.

Tsuchiya Y(1), Kinoshita K, Nakamura H.

Author information:

(1)Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka,

565-0871, Japan.

PreDs is a WWW server that predicts the dsDNA-binding sites on protein molecular

surfaces generated from the atomic coordinates in a PDB format. The prediction

was done by evaluating the electrostatic potential, the local curvature and the

global curvature on the surfaces. Results of the prediction can be interactively

checked with our original surface viewer.AVAILABILITY: PreDs is available free of

charge from http://pre-s.protein.osaka-u.ac.jp/~preds/

CONTACT: kino@ims.u-tokyo.ac.jp.

DOI: 10.1093/bioinformatics/bti232

PMID: 15613393 [Indexed for MEDLINE]

3179. Bioinformatics. 2005 Apr 15;21(8):1376-82. Epub 2004 Dec 10.

Highly specific and accurate selection of siRNAs for high-throughput functional

assays.

Santoyo J(1), Vaquerizas JM, Dopazo J.

Author information:

(1)Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO) and

Functional Genomics Node, INB, Melchor Fernández Almagro, 3, 28029 Madrid, Spain.

MOTIVATION: Small interfering RNA (siRNA) is widely used in functional genomics

to silence genes by decreasing their expression to study the resulting

phenotypes. The possibility of performing large-scale functional assays by gene

silencing accentuates the necessity of a software capable of the high-throughput

design of highly specific siRNA. The main objective sought was the design of a

large number of siRNAs with appropriate thermodynamic properties and, especially,

high specificity. Since all the available procedures require, to some extent,

manual processing of the results to guarantee specific results, specificity

constitutes to date, the major obstacle to the complete automation of all the

steps necessary for the selection of optimal candidate siRNAs.

RESULT: Here, we present a program that for the first time completely automates

the search for siRNAs. In SiDE, the most complete set of rules for the selection

of siRNA candidates (including G+C content, nucleotides at determined positions,

thermodynamic properties, propensity to form internal hairpins, etc.) is

implemented and moreover, specificity is achieved by a conceptually new method.

After selecting possible siRNA candidates with the optimal functional properties,

putative unspecific matches, which can cause cross-hybridization, are checked in

databases containing a unique entry for each gene. These truly non-redundant

databases are constructed from the genome annotations (Ensembl). Also intron/exon

boundaries, presence of polymorphisms (single nucleotide polymorphisms)

specificity for either gene or transcript, and other features can be selected to

be considered in the design of siRNAs.

AVAILABILITY: The program is available as a web server at

http://side.bioinfo.cnio.es. The program was written under the GPL license.

CONTACT: jdopazo@cnio.es.

DOI: 10.1093/bioinformatics/bti196

PMID: 15591357 [Indexed for MEDLINE]

3180. Bioinformatics. 2005 Apr 15;21(8):1719-20. Epub 2004 Dec 7.

Porter: a new, accurate server for protein secondary structure prediction.

Pollastri G(1), McLysaght A.

Author information:

(1)Computer Science Department, University College Dublin, Belfield, Dublin 4,

Ireland. gianluca.pollastri@ucd.ie

Porter is a new system for protein secondary structure prediction in three

classes. Porter relies on bidirectional recurrent neural networks with shortcut

connections, accurate coding of input profiles obtained from multiple sequence

alignments, second stage filtering by recurrent neural networks, incorporation of

long range information and large-scale ensembles of predictors. Porter's

accuracy, tested by rigorous 5-fold cross-validation on a large set of proteins,

exceeds 79%, significantly above a copy of the state-of-the-art SSpro server,

better than any system published to date.AVAILABILITY: Porter is available as a

public web server at http://distill.ucd.ie/porter/

CONTACT: gianluca.pollastri@ucd.ie.

DOI: 10.1093/bioinformatics/bti203

PMID: 15585524 [Indexed for MEDLINE]

3181. Bioinformatics. 2005 Apr 15;21(8):1693-4. Epub 2004 Nov 25.

Procom: a web-based tool to compare multiple eukaryotic proteomes.

Li JB(1), Zhang M, Dutcher SK, Stormo GD.

Author information:

(1)Department of Genetics, Washington University School of Medicine, St. Louis,

MO 63110, USA. billy@ural.wustl.edu

Each organism has traits that are shared with some, but not all, organisms.

Identification of genes needed for a particular trait can be accomplished by a

comparative genomics approach using three or more organisms. Genes that occur in

organisms without the trait are removed from the set of genes in common among

organisms with the trait. To facilitate these comparisons, a web-based server,

Procom, was developed to identify the subset of genes that may be needed for a

trait.AVAILABILITY: The Procom program is freely available with documentation and

examples at http://ural.wustl.edu/~billy/Procom/

CONTACT: billy@ural.wustl.edu.

DOI: 10.1093/bioinformatics/bti161

PMID: 15564299 [Indexed for MEDLINE]

3182. J Biol Chem. 2005 Apr 15;280(15):14427-32. Epub 2005 Jan 12.

Support vector machine-based method for subcellular localization of human

proteins using amino acid compositions, their order, and similarity search.

Garg A(1), Bhasin M, Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39A,

Chandigarh, India.

Here we report a systematic approach for predicting subcellular localization

(cytoplasm, mitochondrial, nuclear, and plasma membrane) of human proteins.

First, support vector machine (SVM)-based modules for predicting subcellular

localization using traditional amino acid and dipeptide (i + 1) composition

achieved overall accuracy of 76.6 and 77.8%, respectively. PSI-BLAST, when

carried out using a similarity-based search against a nonredundant data base of

experimentally annotated proteins, yielded 73.3% accuracy. To gain further

insight, a hybrid module (hybrid1) was developed based on amino acid composition,

dipeptide composition, and similarity information and attained better accuracy of

84.9%. In addition, SVM modules based on a different higher order dipeptide i.e.

i + 2, i + 3, and i + 4 were also constructed for the prediction of subcellular

localization of human proteins, and overall accuracy of 79.7, 77.5, and 77.1% was

accomplished, respectively. Furthermore, another SVM module hybrid2 was developed

using traditional dipeptide (i + 1) and higher order dipeptide (i + 2, i + 3, and

i + 4) compositions, which gave an overall accuracy of 81.3%. We also developed

SVM module hybrid3 based on amino acid composition, traditional and higher order

dipeptide compositions, and PSI-BLAST output and achieved an overall accuracy of

84.4%. A Web server HSLPred (www.imtech.res.in/raghava/hslpred/ or

bioinformatics.uams.edu/raghava/hslpred/) has been designed to predict

subcellular localization of human proteins using the above approaches.

DOI: 10.1074/jbc.M411789200

PMID: 15647269 [Indexed for MEDLINE]

3183. Curr Protoc Bioinformatics. 2005 Apr;Chapter 10:Unit10.4. doi:

10.1002/0471250953.bi1004s9.

MultiPipMaker: comparative alignment server for multiple DNA sequences.

Elnitski L(1), Riemer C, Burhans R, Hardison R, Miller W.

Author information:

(1)The Pennsylvania State University, University Park, Pennsylvania, USA.

The MultiPipMaker World Wide Web server (http://www.bx.psu.edu) provides a useful

tool for aligning multiple sequences and visualizing regions of conservation

between them. This unit describes the use of the MultiPipMaker server and gives

an explanation of the resulting output files and supporting tools. Features

provided by the server include alignment of up to 20 very long genomic sequences,

output choices of a true, nucleotide-level multiple alignment or stacked,

pairwise percent identity plots, and user-specified annotations for genomic

features and elements of choice, with clickable links to additional information.

Alignments can include unordered, unoriented secondary sequences.

DOI: 10.1002/0471250953.bi1004s9

PMID: 18428743 [Indexed for MEDLINE]

3184. Proteins. 2005 Apr 1;59(1):30-7.

Prediction of protein relative solvent accessibility with a two-stage SVM

approach.

Nguyen MN(1), Rajapakse JC.

Author information:

(1)BioInformatics Research Centre, School of Computer Engineering, Nanyang

Technological University, Singapore.

Information on relative solvent accessibility (RSA) of amino acid residues in

proteins provides valuable clues to the prediction of protein structure and

function. A two-stage approach with support vector machines (SVMs) is proposed,

where an SVM predictor is introduced to the output of the single-stage SVM

approach to take into account the contextual relationships among solvent

accessibilities for the prediction. By using the position-specific scoring

matrices (PSSMs) generated by PSI-BLAST, the two-stage SVM approach achieves

accuracies up to 90.4% and 90.2% on the Manesh data set of 215 protein structures

and the RS126 data set of 126 nonhomologous globular proteins, respectively,

which are better than the highest published scores on both data sets to date. A

Web server for protein RSA prediction using a two-stage SVM method has been

developed and is available (http://birc.ntu.edu.sg/~pas0186457/rsa.html).

2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20404

PMID: 15696542 [Indexed for MEDLINE]

3185. Rofo. 2005 Apr;177(4):569-75.

[Quality improvement of resources in radiology on the internet].

[Article in German]

Grunewald M(1), Gebhard H, Wagner M, Bautz WA, Alibek S.

Author information:

(1)Institut für Diagnostische Radiologie, Friedrich-Alexander-Universität

Erlangen-Nürnberg. markus.grunewald@idr.imed.uni-erlangen.de

PURPOSE: Categorization and evaluation of online teaching files in radiology by

representative members of the target group to make the specific search for

adequate programs more effective.

MATERIALS AND METHODS: A representative team of board qualified radiologists,

residents and medical students performed a basic search for radiology teaching

files on the Internet using search machines, international mailing lists and link

lists to collections of national and international radiological societies and

departments. The programs were categorized by language, modality, target group

and special features, such as qualification for CME-accreditation. For final

evaluation and ranking of the detected files, a questionnaire was developed to

assess completeness, image quality, page loading time, layout, orientation,

interactivity, annotation and maintenance. The results were stored in an Access

database on a web server. A query form in HTML format, including the parameters

described above, was made accessible to the online user.

RESULTS: A search machine for radiological teaching files (RadList/Entity-link

List) was made available online ( www.tnt-radiology.de/radlist and

www.tnt-radiology.de/entitylinklist ). A submitted request calls a cgi script

that searches the database for the appropriate sites according to the individual

search parameters selected by the user. The list of matching URLs is returned to

the user as HTML page. Evaluating the single sites by applying the criteria

listed above contributed to the quality assurance of the radiological teaching

resources on the Internet.

CONCLUSION: Adapting a new Internet interface to the particular needs of the user

allows a more effective access to specific radiological teaching files online.

RadList/Entity-link List ( www.tnt-radiology.de/radlist and

www.tnt-radiology.de/entitylinklist ) is conducive to quality improvement and

benefits users as well as authors of radiological teaching files on the Internet.

DOI: 10.1055/s-2005-857904

PMID: 15838764 [Indexed for MEDLINE]

3186. BMC Bioinformatics. 2005 Mar 17;6:59.

Correlation and prediction of gene expression level from amino acid and dipeptide

composition of its protein.

Raghava GP(1), Han JH.

Author information:

(1)Department of Computer Science and Engineering, Pohang University of Science

and Technology, San 31 Hyo-Ja Dong, Pohang 790-784, Republic of Korea.

raghava@imtech.res.in

BACKGROUND: A large number of papers have been published on analysis of

microarray data with particular emphasis on normalization of data, detection of

differentially expressed genes, clustering of genes and regulatory network. On

other hand there are only few studies on relation between expression level and

composition of nucleotide/protein sequence, using expression data. There is a

need to understand why particular genes/proteins express more in particular

conditions. In this study, we analyze 3468 genes of Saccharomyces cerevisiae

obtained from Holstege et al., (1998) to understand the relationship between

expression level and amino acid composition.

RESULTS: We compute the correlation between expression of a gene and amino acid

composition of its protein. It was observed that some residues (like Ala, Gly,

Arg and Val) have significant positive correlation (r > 0.20) and some other

residues (Like Asp, Leu, Asn and Ser) have negative correlation (r < -0.15) with

the expression of genes. A significant negative correlation (r = -0.18) was also

found between length and gene expression. These observations indicate the

relationship between percent composition and gene expression level. Thus,

attempts have been made to develop a Support Vector Machine (SVM) based method

for predicting the expression level of genes from its protein sequence. In this

method the SVM is trained with proteins whose gene expression data is known in a

given condition. Then trained SVM is used to predict the gene expression of other

proteins of the same organism in the same condition. A correlation coefficient r

= 0.70 was obtained between predicted and experimentally determined expression of

genes, which improves from r = 0.70 to 0.72 when dipeptide composition was used

instead of residue composition. The method was evaluated using 5-fold cross

validation test. We also demonstrate that amino acid composition information

along with gene expression data can be used for improving the function

classification of proteins.

CONCLUSION: There is a correlation between gene expression and amino acid

composition that can be used to predict the expression level of genes up to a

certain extent. A web server based on the above strategy has been developed for

calculating the correlation between amino acid composition and gene expression

and prediction of expression level http://kiwi.postech.ac.kr/raghava/lgepred/.

This server will allow users to study the evolution from expression data.

DOI: 10.1186/1471-2105-6-59

PMCID: PMC1083413

PMID: 15773999 [Indexed for MEDLINE]

3187. BMC Bioinformatics. 2005 Mar 11;6:52.

WEBnm@: a web application for normal mode analyses of proteins.

Hollup SM(1), Salensminde G, Reuter N.

Author information:

(1)Computational Biology Unit, Bergen Center for Computational Science,

University of Bergen, Thormøhlensgt. 55, N-5008 Bergen, Norway. sivh@ii.uib.no

BACKGROUND: Normal mode analysis (NMA) has become the method of choice to

investigate the slowest motions in macromolecular systems. NMA is especially

useful for large biomolecular assemblies, such as transmembrane channels or virus

capsids. NMA relies on the hypothesis that the vibrational normal modes having

the lowest frequencies (also named soft modes) describe the largest movements in

a protein and are the ones that are functionally relevant.

RESULTS: We developed a web-based server to perform normal modes calculations and

different types of analyses. Starting from a structure file provided by the user

in the PDB format, the server calculates the normal modes and subsequently offers

the user a series of automated calculations; normalized squared atomic

displacements, vector field representation and animation of the first six

vibrational modes. Each analysis is performed independently from the others and

results can be visualized using only a web browser. No additional plug-in or

software is required. For users who would like to analyze the results with their

favorite software, raw results can also be downloaded. The application is

available on http://www.bioinfo.no/tools/normalmodes. We present here the

underlying theory, the application architecture and an illustration of its

features using a large transmembrane protein as an example.

CONCLUSION: We built an efficient and modular web application for normal mode

analysis of proteins. Non specialists can easily and rapidly evaluate the degree

of flexibility of multi-domain protein assemblies and characterize the large

amplitude movements of their domains.

DOI: 10.1186/1471-2105-6-52

PMCID: PMC1274249

PMID: 15762993 [Indexed for MEDLINE]

3188. BMC Bioinformatics. 2005 Mar 2;6:41.

PHACCS, an online tool for estimating the structure and diversity of uncultured

viral communities using metagenomic information.

Angly F(1), Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P,

Felts B, Nulton J, Mahaffy J, Rohwer F.

Author information:

(1)Ecole Supérieure de Biotechnologie de Strasbourg, Boulevard Sébastien Brandt,

67413 Illkirch, France. fangly@projects.sdsu.edu <fangly@projects.sdsu.edu>

BACKGROUND: Phages, viruses that infect prokaryotes, are the most abundant

microbes in the world. A major limitation to studying these viruses is the

difficulty of cultivating the appropriate prokaryotic hosts. One way around this

limitation is to directly clone and sequence shotgun libraries of uncultured

viral communities (i.e., metagenomic analyses). PHACCS

http://phage.sdsu.edu/phaccs, Phage Communities from Contig Spectrum, is an

online bioinformatic tool to assess the biodiversity of uncultured viral

communities. PHACCS uses the contig spectrum from shotgun DNA sequence assemblies

to mathematically model the structure of viral communities and make predictions

about diversity.

RESULTS: PHACCS builds models of possible community structure using a modified

Lander-Waterman algorithm to predict the underlying contig spectrum. PHACCS finds

the most appropriate structure model by optimizing the model parameters until the

predicted contig spectrum is as close as possible to the experimental one. This

model is the basis for making estimates of uncultured viral community richness,

evenness, diversity index and abundance of the most abundant genotype.

CONCLUSION: PHACCS analysis of four different environmental phage communities

suggests that the power law is an important rank-abundance form to describe

uncultured viral community structure. The estimates support the fact that the

four phage communities were extremely diverse and that phage community

biodiversity and structure may be correlated with that of their hosts.

DOI: 10.1186/1471-2105-6-41

PMCID: PMC555943

PMID: 15743531 [Indexed for MEDLINE]

3189. Biophysics (Nagoya-shi). 2005 Feb 28;1:21-24. eCollection 2005.

HINT: a database of annotated protein-protein interactions and their homologs.

Patil A(1), Nakamura H(1).

Author information:

(1)Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka

565-0871, Japan.

Despite the abundance of protein-protein interaction databases currently

available online, a source that identifies and lists similar interactions in

different species is lacking. The Homologous Interactions (HINT) database is such

a collection of protein-protein interactions and their homologs in one or more

species. The interactions and their homologs are annotated with Eukaryotic

Cluster of Orthologous Groups (KOG) IDs, InterPro domains, Gene Ontology (GO)

terminology and Protein Data Bank (PDB) structures. HINT is available as an

interactive Web server at http://helix.protein.osaka-u.ac.jp/hint/.

DOI: 10.2142/biophysics.1.21

PMCID: PMC5036632

PMID: 27857549

3190. BMC Genomics. 2005 Feb 21;6:24.

Comparative promoter region analysis powered by CORG.

Dieterich C(1), Grossmann S, Tanzer A, Röpcke S, Arndt PF, Stadler PF, Vingron M.

Author information:

(1)Computational Molecular Biology Department, Max Planck Institute for Molecular

Genetics, Ihnestrasse 73, 14195 Berlin, Germany.

christoph.dieterich@molgen.mpg.de

BACKGROUND: Promoters are key players in gene regulation. They receive signals

from various sources (e.g. cell surface receptors) and control the level of

transcription initiation, which largely determines gene expression. In

vertebrates, transcription start sites and surrounding regulatory elements are

often poorly defined. To support promoter analysis, we present CORG

http://corg.molgen.mpg.de, a framework for studying upstream regions including

untranslated exons (5' UTR).

DESCRIPTION: The automated annotation of promoter regions integrates information

of two kinds. First, statistically significant cross-species conservation within

upstream regions of orthologous genes is detected. Pairwise as well as multiple

sequence comparisons are computed. Second, binding site descriptions

(position-weight matrices) are employed to predict conserved regulatory elements

with a novel approach. Assembled EST sequences and verified transcription start

sites are incorporated to distinguish exonic from other sequences. As of now, we

have included 5 species in our analysis pipeline (man, mouse, rat, fugu and

zebrafish). We characterized promoter regions of 16,127 groups of orthologous

genes. All data are presented in an intuitive way via our web site. Users are

free to export data for single genes or access larger data sets via our DAS

server http://tomcat.molgen.mpg.de:8080/das. The benefits of our framework are

exemplarily shown in the context of phylogenetic profiling of transcription

factor binding sites and detection of microRNAs close to transcription start

sites of our gene set.

CONCLUSION: The CORG platform is a versatile tool to support analyses of gene

regulation in vertebrate promoter regions. Applications for CORG cover a broad

range from studying evolution of DNA binding sites and promoter constitution to

the discovery of new regulatory sequence elements (e.g. microRNAs and binding

sites).

DOI: 10.1186/1471-2164-6-24

PMCID: PMC555765

PMID: 15723697 [Indexed for MEDLINE]

3191. BMC Bioinformatics. 2005 Feb 19;6:33.

PSSM-based prediction of DNA binding sites in proteins.

Ahmad S(1), Sarai A.

Author information:

(1)Department of Bioinformatics and Bioscience, Kyushu Institute of Technology,

Iizuka 820 8502, Fukuoka, Japan. shandar@bse.kyutech.ac.jp

BACKGROUND: Detection of DNA-binding sites in proteins is of enormous interest

for technologies targeting gene regulation and manipulation. We have previously

shown that a residue and its sequence neighbor information can be used to predict

DNA-binding candidates in a protein sequence. This sequence-based prediction

method is applicable even if no sequence homology with a previously known

DNA-binding protein is observed. Here we implement a neural network based

algorithm to utilize evolutionary information of amino acid sequences in terms of

their position specific scoring matrices (PSSMs) for a better prediction of

DNA-binding sites.

RESULTS: An average of sensitivity and specificity using PSSMs is up to 8.7%

better than the prediction with sequence information only. Much smaller data sets

could be used to generate PSSM with minimal loss of prediction accuracy.

CONCLUSION: One problem in using PSSM-derived prediction is obtaining lengthy and

time-consuming alignments against large sequence databases. In order to speed up

the process of generating PSSMs, we tried to use different reference data sets

(sequence space) against which a target protein is scanned for PSI-BLAST

iterations. We find that a very small set of proteins can actually be used as

such a reference data without losing much of the prediction value. This makes the

process of generating PSSMs very rapid and even amenable to be used at a genome

level. A web server has been developed to provide these predictions of

DNA-binding sites for any new protein from its amino acid sequence.

AVAILABILITY: Online predictions based on this method are available at

http://www.netasa.org/dbs-pssm/

DOI: 10.1186/1471-2105-6-33

PMCID: PMC550660

PMID: 15720719 [Indexed for MEDLINE]

3192. Bioinformatics. 2005 Feb 15;21(4):557-9. Epub 2004 Sep 16.

MAP-O-MAT: internet-based linkage mapping.

Kong X(1), Matise TC.

Author information:

(1)Department of Genetics, Rutgers University Piscataway NJ 08854, USA.

MAP-O-MAT is a web-based server for automated linkage mapping of human

polymorphic DNA markers. MAP-O-MAT facilitates the verification of order and map

distances for custom mapping sets using genotype data from the CEPH database, and

from the Marshfield, SNP Consortium and Rutgers linkage maps (exclusive to the

deCODE genotyping data). The CRI-MAP program is used for likelihood calculations

and some mapping algorithms, and physical map positions are provided from the

human genome assembly.AVAILABILITY: MAP-O-MAT is located at

http://compgen.rutgers.edu/mapomat/

CONTACT: matise@biology.rutgers.edu.

DOI: 10.1093/bioinformatics/bti024

PMID: 15374870 [Indexed for MEDLINE]

3193. Bioinformatics. 2005 Feb 15;21(4):545-7. Epub 2004 Sep 16.

Online synonymous codon usage analyses with the ade4 and seqinR packages.

Charif D(1), Thioulouse J, Lobry JR, Perrière G.

Author information:

(1)Laboratoire de Biométrie et Biologie Evolutive-CNRS UMR 5558, and INRIA Helix

project, Université Claude Bernard-Lyon I 43 bd. du 11 Novembre 1918, F-69622

Villeurbanne cedex, France.

Correspondence analysis of codon usage data is a widely used method in sequence

analysis, but the variability in amino acid composition between proteins is a

confounding factor when one wants to analyse synonymous codon usage variability.

A simple and natural way to cope with this problem is to use within-group

correspondence analysis. There is, however, no user-friendly implementation of

this method available for genomic studies. Our motivation was to provide to the

community a Web facility to easily study synonymous codon usage on a subset of

data available in public genomic databases.AVAILABILITY: Availability through the

Pole Bioinformatique Lyonnais (PBIL) Web server at

http://pbil.univ-lyon1.fr/datasets/charif04/ with a demo allowing us to reproduce

the figure in the present application note. All underlying software is

distributed under a GPL licence.

CONTACT: http://pbil.univ-lyon1.fr/members/lobry.

DOI: 10.1093/bioinformatics/bti037

PMID: 15374859 [Indexed for MEDLINE]

3194. Proteins. 2005 Feb 15;58(3):610-7.

The ConSurf-HSSP database: the mapping of evolutionary conservation among

homologs onto PDB structures.

Glaser F(1), Rosenberg Y, Kessel A, Pupko T, Ben-Tal N.

Author information:

(1)Department of Biochemistry, Tel Aviv University, Tel Aviv, Israel.

The HSSP (Homology-Derived Secondary Structure of Proteins) database provides

multiple sequence alignments (MSAs) for proteins of known three-dimensional (3D)

structure in the Protein Data Bank (PDB). The database also contains an estimate

of the degree of evolutionary conservation at each amino acid position. This

estimate, which is based on the relative entropy, correlates with the functional

importance of the position; evolutionarily conserved positions (i.e., positions

with limited variability and low entropy) are occasionally important to maintain

the 3D structure and biological function(s) of the protein. We recently developed

the Rate4Site algorithm for scoring amino acid conservation based on their

calculated evolutionary rate. This algorithm takes into account the phylogenetic

relationships between the homologs and the stochastic nature of the evolutionary

process. Here we present the ConSurf-HSSP database of Rate4Site estimates of the

evolutionary rates of the amino acid positions, calculated using HSSP's MSAs. The

database provides precalculated evolutionary rates for nearly all of the PDB.

These rates are projected, using a color code, onto the protein structure, and

can be viewed online using the ConSurf server interface. To exemplify the

database, we analyzed in detail the conservation pattern obtained for pyruvate

kinase and compared the results with those observed using the relative entropy

scores of the HSSP database. It is reassuring to know that the main functional

region of the enzyme is detectable using both conservation scores. Interestingly,

the ConSurf-HSSP calculations mapped additional functionally important regions,

which are moderately conserved and were overlooked by the original HSSP estimate.

The ConSurf-HSSP database is available online (http://consurf-hssp.tau.ac.il).

(c) 2004 Wiley-Liss, Inc.

DOI: 10.1002/prot.20305

PMID: 15614759 [Indexed for MEDLINE]

3195. Proteins. 2005 Feb 15;58(3):618-27.

FAST: a novel protein structure alignment algorithm.

Zhu J(1), Weng Z.

Author information:

(1)Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA.

We present a novel algorithm named FAST for aligning protein three-dimensional

structures. FAST uses a directionality-based scoring scheme to compare the

intra-molecular residue-residue relationships in two structures. It employs an

elimination heuristic to promote sparseness in the residue-pair graph and

facilitate the detection of the global optimum. In order to test the overall

accuracy of FAST, we determined its sensitivity and specificity with the SCOP

classification (version 1.61) as the gold standard. FAST achieved higher

sensitivities than several existing methods (DaliLite, CE, and K2) at all

specificity levels. We also tested FAST against 1033 manually curated alignments

in the HOMSTRAD database. The overall agreement was 96%. Close inspection of

examples from broad structural classes indicated the high quality of FAST

alignments. Moreover, FAST is an order of magnitude faster than other algorithms

that attempt to establish residue-residue correspondence. Typical pairwise

alignments take FAST less than a second with a Pentium III 1.2GHz CPU. FAST

software and a web server are available at http://biowulf.bu.edu/FAST/.

(c) 2004 Wiley-Liss, Inc.

DOI: 10.1002/prot.20331

PMID: 15609341 [Indexed for MEDLINE]

3196. Bioinformatics. 2005 Feb 1;21(3):393-5. Epub 2004 Sep 3.

The SSEA server for protein secondary structure alignment.

Fontana P(1), Bindewald E, Toppo S, Velasco R, Valle G, Tosatto SC.

Author information:

(1)Istituto Agrario di San Michele all'Adige, via E. Mach 1, 38010 S. Michele

all'Adige (TN), Italy.

SUMMARY: We present a web server that computes alignments of protein secondary

structures. The server supports both performing pairwise alignments and searching

a secondary structure against a library of domain folds. It can calculate global

and local secondary structure element alignments. A combination of local and

global alignment steps can be used to search for domains inside the query

sequence or help in the discrimination of novel folds. Both the SCOP and PDB fold

libraries, clustered at 95 and 40% sequence identity, are available for

alignment.

AVAILABILITY: The web server interface is freely accessible to academic users at

http://protein.cribi.unipd.it/ssea/. The executable version and benchmarking data

are available from the same web page.

DOI: 10.1093/bioinformatics/bti013

PMID: 15347578 [Indexed for MEDLINE]

3197. Bioinformatics. 2005 Feb 1;21(3):388-9. Epub 2004 Sep 3.

Tracker: continuous HMMER and BLAST searching.

Marchin M(1), Kelly PT, Fang J.

Author information:

(1)Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66045,

USA.

SUMMARY: Tracker is a web-based email alert system for monitoring protein

database searches using HMMER and Blast-P, nucleotide searches using Blast-N and

literature searches of the PubMed database. Users submit searches via a web-based

interface. Searches are saved and run against updated databases to alert users

about new information. If there are new results from the saved searches, users

will be notified by email and will then be able to access results and link to

additional information on the NCBI website. Tracker supports Boolean AND/OR

operations on HMMER and BLASTP result sets to allow users to broaden or narrow

protein searches.

AVAILABILITY: The server is located at

http://jay.bioinformatics.ku.edu/tracker/index.html. A distribution package

including detailed installation procedure is freely available from

http://jay.bioinformatics.ku.edu/download/tracker/.

DOI: 10.1093/bioinformatics/bti012

PMID: 15347577 [Indexed for MEDLINE]

3198. BMC Bioinformatics. 2005 Feb 1;6:20.

GeneNotes--a novel information management software for biologists.

Hong P(1), Wong WH.

Author information:

(1)BioX program, Department of Statistics and HRP, Stanford University, Stanford,

CA 94305-4065, USA. pengyuhong@stanford.edu

BACKGROUND: Collecting and managing information is a challenging task in a

genome-wide profiling research project. Most databases and online computational

tools require a direct human involvement. Information and computational results

are presented in various multimedia formats (e.g., text, image, PDF, word files,

etc.), many of which cannot be automatically processed by computers in

biologically meaningful ways. In addition, the quality of computational results

is far from perfect and requires nontrivial manual examination. The timely

selection, integration and interpretation of heterogeneous biological information

still heavily rely on the sensibility of biologists. Biologists often feel

overwhelmed by the huge amount of and the great diversity of distributed

heterogeneous biological information.

DESCRIPTION: We developed an information management application called GeneNotes.

GeneNotes is the first application that allows users to collect and manage

multimedia biological information about genes/ESTs. GeneNotes provides an

integrated environment for users to surf the Internet, collect notes for

genes/ESTs, and retrieve notes. GeneNotes is supported by a server that

integrates gene annotations from many major databases (e.g., HGNC, MGI, etc.).

GeneNotes uses the integrated gene annotations to (a) identify genes given

various types of gene IDs (e.g., RefSeq ID, GenBank ID, etc.), and (b) provide

quick views of genes. GeneNotes is free for academic usage. The program and the

tutorials are available at: http://bayes.fas.harvard.edu/genenotes/.

CONCLUSIONS: GeneNotes provides a novel human-computer interface to assist

researchers to collect and manage biological information. It also provides a

platform for studying how users behave when they manipulate biological

information. The results of such study can lead to innovation of more intelligent

human-computer interfaces that greatly shorten the cycle of biology research.

DOI: 10.1186/1471-2105-6-20

PMCID: PMC549201

PMID: 15686593 [Indexed for MEDLINE]

3199. J Bioinform Comput Biol. 2005 Feb;3(1):35-60.

Optimizing long intrinsic disorder predictors with protein evolutionary

information.

Peng K(1), Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z.

Author information:

(1)Center for Information Science and Technology, Temple University,

Philadelphia, PA 19122, USA.

Protein existing as an ensemble of structures, called intrinsically disordered,

has been shown to be responsible for a wide variety of biological functions and

to be common in nature. Here we focus on improving sequence-based predictions of

long (>30 amino acid residues) regions lacking specific 3-D structure by means of

four new neural-network-based Predictors Of Natural Disordered Regions (PONDRs):

VL3, VL3H, VL3P, and VL3E. PONDR VL3 used several features from a previously

introduced PONDR VL2, but benefitted from optimized predictor models and a

slightly larger (152 vs. 145) set of disordered proteins that were cleaned of

mislabeling errors found in the smaller set. PONDR VL3H utilized homologues of

the disordered proteins in the training stage, while PONDR VL3P used attributes

derived from sequence profiles obtained by PSI-BLAST searches. The measure of

accuracy was the average between accuracies on disordered and ordered protein

regions. By this measure, the 30-fold cross-validation accuracies of VL3, VL3H,

and VL3P were, respectively, 83.6 +/- 1.4%, 85.3 +/- 1.4%, and 85.2 +/- 1.5%. By

combining VL3H and VL3P, the resulting PONDR VL3E achieved an accuracy of 86.7

+/- 1.4%. This is a significant improvement over our previous PONDRs VLXT (71.6

+/- 1.3%) and VL2 (80.9 +/- 1.4%). The new disorder predictors with the

corresponding datasets are freely accessible through the web server at

http://www.ist.temple.edu/disprot.

PMID: 15751111 [Indexed for MEDLINE]

3200. Proteins. 2005 Feb 1;58(2):321-8.

Fold recognition by combining sequence profiles derived from evolution and from

depth-dependent structural alignment of fragments.

Zhou H(1), Zhou Y.

Author information:

(1)Howard Hughes Medical Institute Center for Single Molecule Biophysics,

Department of Physiology & Biophysics, State University of New York at Buffalo,

14214, USA.

Recognizing structural similarity without significant sequence identity has

proved to be a challenging task. Sequence-based and structure-based methods as

well as their combinations have been developed. Here, we propose a

fold-recognition method that incorporates structural information without the need

of sequence-to-structure threading. This is accomplished by generating sequence

profiles from protein structural fragments. The structure-derived sequence

profiles allow a simple integration with evolution-derived sequence profiles and

secondary-structural information for an optimized alignment by efficient dynamic

programming. The resulting method (called SP(3)) is found to make a statistically

significant improvement in both sensitivity of fold recognition and accuracy of

alignment over the method based on evolution-derived sequence profiles alone (SP)

and the method based on evolution-derived sequence profile and secondary

structure profile (SP(2)). SP(3) was tested in SALIGN benchmark for alignment

accuracy and Lindahl, PROSPECTOR 3.0, and LiveBench 8.0 benchmarks for

remote-homology detection and model accuracy. SP(3) is found to be the most

sensitive and accurate single-method server in all benchmarks tested where other

methods are available for comparison (although its results are statistically

indistinguishable from the next best in some cases and the comparison is

subjected to the limitation of time-dependent sequence and/or structural library

used by different methods.). In LiveBench 8.0, its accuracy rivals some of the

consensus methods such as ShotGun-INBGU, Pmodeller3, Pcons4, and ROBETTA. SP(3)

fold-recognition server is available on http://theory.med.buffalo.edu.

(c) 2004 Wiley-Liss, Inc.

DOI: 10.1002/prot.20308

PMCID: PMC1408319

PMID: 15523666 [Indexed for MEDLINE]

3201. Nucleic Acids Res. 2005 Jan 28;33(2):616-21. Print 2005.

Measuring genome conservation across taxa: divided strains and united kingdoms.

Kunin V(1), Ahren D, Goldovsky L, Janssen P, Ouzounis CA.

Author information:

(1)Computational Genomics Group, The European Bioinformatics Institute EMBL

Cambridge Outstation, Cambridge CB10 1SD, UK.

Species evolutionary relationships have traditionally been defined by sequence

similarities of phylogenetic marker molecules, recently followed by whole-genome

phylogenies based on gene order, average ortholog similarity or gene content.

Here, we introduce genome conservation--a novel metric of evolutionary distances

between species that simultaneously takes into account, both gene content and

sequence similarity at the whole-genome level. Genome conservation represents a

robust distance measure, as demonstrated by accurate phylogenetic

reconstructions. The genome conservation matrix for all presently sequenced

organisms exhibits a remarkable ability to define evolutionary relationships

across all taxonomic ranges. An assessment of taxonomic ranks with genome

conservation shows that certain ranks are inadequately described and raises the

possibility for a more precise and quantitative taxonomy in the future. All

phylogenetic reconstructions are available at the genome phylogeny server:

<http://maine.ebi.ac.uk:8000/cgi-bin/gps/GPS.pl>.

DOI: 10.1093/nar/gki181

PMCID: PMC548337

PMID: 15681613 [Indexed for MEDLINE]

3202. Int J Health Geogr. 2005 Jan 13;4(1):2.

Research protocol: EB-GIS4HEALTH UK - foundation evidence base and ontology-based

framework of modular, reusable models for UK/NHS health and healthcare GIS

applications.

Boulos MN(1).

Author information:

(1)School for Health, University of Bath, Claverton Down, Bath BA2 7AY, UK.

M.N.K.Boulos@bath.ac.uk.

EB-GIS4HEALTH UK aims at building a UK-oriented foundation evidence base and

modular conceptual models for GIS applications and programmes in health and

healthcare to improve the currently poor GIS state of affairs within the NHS;

help the NHS understand and harness the importance of spatial information in the

health sector in order to better respond to national health plans, priorities,

and requirements; and also foster the much-needed NHS-academia GIS collaboration.

The project will focus on diabetes and dental care, which together account for

about 11% of the annual NHS budget, and are thus important topics where GIS can

help optimising resource utilisation and outcomes. Virtual e-focus groups will

ensure all UK/NHS health GIS stakeholders are represented. The models will be

built using Protege ontology editor http://protege.stanford.edu/ based on the

best evidence pooled in the project's evidence base (from critical literature

reviews and e-focus groups). We will disseminate our evidence base, GIS models,

and documentation through the project's Web server. The models will be

human-readable in different ways to inform NHS GIS implementers, and it will be

possible to also use them to generate the necessary template databases (and even

to develop "intelligent" health GIS solutions using software agents) for running

the modelled applications. Our products and experience in this project will be

transferable to address other national health topics based on the same

principles. Our ultimate goal is to provide the NHS with practical,

vendor-neutral, modular workflow models, and ready-to-use, evidence-based

frameworks for developing successful GIS business plans and implementing GIS to

address various health issues. NHS organisations adopting such frameworks will

achieve a common understanding of spatial data and processes, which will enable

them to efficiently and effectively share, compare, and integrate their data

silos and results for more informed planning and better outcomes.

DOI: 10.1186/1476-072X-4-2

PMCID: PMC546191

PMID: 15649328

3203. Appl Bioinformatics. 2005;4(2):147-50.

ProLysED: an integrated database and meta-server of bacterial protease systems.

Firdaus Raih M(1), Ahmad HA, Sharum MY, Azizi N, Mohamed R.

Author information:

(1)School of BioSciences and Biotechnology, Faculty of Science and Technology,

Universiti Kebangsaan Malaysia, Bangi, Malaysia. mfirr@pkrisc.cc.ukm.my

Bacterial proteases are an important group of enzymes that have very diverse

biochemical and cellular functions. Proteases from prokaryotic sources also have

a wide range of uses, either in medicine as pathogenic factors or in industry and

therapeutics. ProLysED (Prokaryotic Lysis Enzymes Database), our meta-server

integrated database of bacterial proteases, is a useful, albeit very niche,

resource. The features include protease classification browsing and searching,

organism-specific protease browsing, molecular information and visualisation of

protease structures from the Protein Data Bank (PDB) as well as predicted

protease structures.AVAILABILITY: ProLysED is integrated into the ProLysES

(Prokaryotic Lysis Enzymes Site) website at http://genome.ukm.my/prolyses/.

Access to the ProLysED database is free for academic users upon registration.

PMID: 16128617 [Indexed for MEDLINE]

3204. Appl Bioinformatics. 2005;4(2):141-5.

PSST-2.0: Protein Data Bank Sequence Search Tool.

Ananthalakshmi P(1), Samayamohan K, Chokalingam C, Mayilarasi C, Sekar K.

Author information:

(1)Bioinformatics Centre, Indian Institute of Science, Bangalore, India.

PSST-2.0 (Protein Data Bank [PDB] Sequence Search Tool) is an updated version of

the earlier PSST (Protein Sequence Search Tool), and the philosophy behind the

search engine has remained unchanged. PSST-2.0 is a Web-based, interactive search

engine developed to retrieve required protein or nucleic acid sequence

information and some of its related details, primarily from sequences derived

from the structures deposited in the PDB (the database of 3-dimensional [3-D]

protein and nucleic acid structures). Additionally, the search engine works for a

selected subset of 25% or 90% non-homologous protein chains. For some of the

selected options, the search engine produces a detailed output for the

user-uploaded, 3-D atomic coordinates of the protein structure (PDB file format)

from the client machine through the Web browser. The search engine works on a

locally maintained PDB, which is updated every week from the parent server at the

Research Collaboratory for Structural Bioinformatics, and hence the search

results are up to date at any given time.AVAILABILITY: PSST-2.0 is freely

accessible via http://pranag.physics.iisc.ernet.in/psst/ or

http://144.16.71.10/psst/.

PMID: 16128616 [Indexed for MEDLINE]

3205. Bull Hosp Jt Dis. 2005;63(1-2):15-9.

The Swiss Orthopaedic Registry.

Röder C(1), El-Kerdi A, Frigg A, Kolling C, Staub LP, Bach B, Müller U.

Author information:

(1)Institute for Evaluative Research in Orthopaedic Surgery, University of Bern,

Switzerland.

Following the tradition of the IDES European Hip Registry inaugurated by M. E.

Müller in the 1960s, the Institute for Evaluative Research in Orthopaedic Surgery

at the University of Bern started a new era of data collection using internet

technology (www.memdoc.org). With support of the Swiss Orthopaedic Society, the

pilot of the Swiss Orthopaedic Registry was conducted, and in cooperation with

different academic and non-academic centers the practicability of integrating the

various data collection instruments into the daily clinical workflow was

evaluated. Three different sizes of hip and knee questionnaires were compiled,

covering the individual demands of the participating hospitals whereby the

smaller questionnaires always represent a subset of the next larger one.

Different types of data collection instruments are available: the online

interface, optical mark reader paper questionnaires, and barcode sheets. Precise

implant tracking is implemented by scanning the implant barcodes directly in the

operating theaters and linking them to the clinical data set via a central

server. In addition, radiographic information can be linked with the clinical

data set. The pilot clinics suggested enhancements to the user interface and

additional features for data management. Also, recommendations were made to

simplify content in some instances and diversify in others. With a new software

release and adapted questionnaires the Swiss Orthopaedic Registry was officially

launched in Summer 2005.

PMID: 16536212 [Indexed for MEDLINE]

3206. Conf Proc IEEE Eng Med Biol Soc. 2005;3:2599-602.

UMass Morph Server: Macromolecular Dynamics Analyses Using Elastic Network

Models.

Jang Y(1), Kim M.

Author information:

(1)Department of Mechanical and Industrial Engineering, University of

Massachusetts, Amherst, MA 01003 USA.

The geometry-based mechanical models called elastic network models (ENMs) in

various resolutions have been developed for the study of macromolecular motions.

In a coarse-grained ENM, a biological system is represented as a network of

springs connecting representative points. They range from single atoms to

functional domains depending on the level of details in modeling. In this paper

presented are the various kinds of coarse-graining methods such as

symmetry-constrained, rigid-cluster, and hybrid ENMs. They enable us to overcome

the computational burden and memory limitation in the conventional molecular

dynamics (MD) simulations and full-atom normal mode analysis (NMA) without loss

of generality. For the broad impact of this work on the structural biology area

we also develop the UMass Morph Server (UMMS). Based on the requests from online

users, UMMS does not only serve a harmonic NMA that describes thermal behaviors

(i.e., fluctuations) of a macromolecule around its equilibrium state, but also

generates anharmonic transition pathways between two end conformations by using

the elastic network interpolation (ENI) also developed by the author. In

addition, UMMS can provide two unique features as follows: (i) interpretation of

massive MD data by finding essential pathways (ii) the conformation prediction

incorporated with time-resolved information such as FRET data. Many example

movies and numeric data can be downloadable at

http://biomechanics.ecs.umass.edu/umms.html.

DOI: 10.1109/IEMBS.2005.1617001

PMID: 17282770

3207. Genome Biol. 2005;6(1):R2. Epub 2004 Dec 22.

Computational prediction of human metabolic pathways from the complete human

genome.

Romero P(1), Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD.

Author information:

(1)Bioinformatics Research Group, SRI International, 333 Ravenswood Ave, Menlo

Park, CA 94025, USA. promero@ai.sri.com <promero@ai.sri.com>

BACKGROUND: We present a computational pathway analysis of the human genome that

assigns enzymes encoded therein to predicted metabolic pathways. Pathway

assignments place genes in their larger biological context, and are a necessary

first step toward quantitative modeling of metabolism.

RESULTS: Our analysis assigns 2,709 human enzymes to 896 bioreactions; 622 of the

enzymes are assigned roles in 135 predicted metabolic pathways. The predicted

pathways closely match the known nutritional requirements of humans. This

analysis identifies probable omissions in the human genome annotation in the form

of 203 pathway holes (missing enzymes within the predicted pathways). We have

identified putative genes to fill 25 of these holes. The predicted human

metabolic map is described by a Pathway/Genome Database called HumanCyc, which is

available at http://HumanCyc.org/. We describe the generation of HumanCyc, and

present an analysis of the human metabolic map. For example, we compare the

predicted human metabolic pathway complement to the pathways of Escherichia coli

and Arabidopsis thaliana and identify 35 pathways that are shared among all three

organisms.

CONCLUSIONS: Our analysis elucidates a significant portion of the human metabolic

map, and also indicates probable unidentified genes in the genome. HumanCyc

provides a genome-based view of human nutrition that associates the essential

dietary requirements of humans with a set of metabolic pathways whose existence

is supported by the human genome. The database places many human genes in a

pathway context, thereby facilitating analysis of gene expression, proteomics,

and metabolomics datasets through a publicly available online tool called the

Omics Viewer.

DOI: 10.1186/gb-2004-6-1-r2

PMCID: PMC549063

PMID: 15642094 [Indexed for MEDLINE]

3208. Genome Res. 2005 Jan;15(1):184-94. Epub 2004 Dec 8.

Mulan: multiple-sequence local alignment and visualization for studying function

and evolution.

Ovcharenko I(1), Loots GG, Giardine BM, Hou M, Ma J, Hardison RC, Stubbs L,

Miller W.

Author information:

(1)Energy, Environment, Biology and Institutional Computing, Lawrence Livermore

National Laboratory, Livermore, California 94550, USA. ovcharenko1@llnl.gov

Multiple-sequence alignment analysis is a powerful approach for understanding

phylogenetic relationships, annotating genes, and detecting functional regulatory

elements. With a growing number of partly or fully sequenced vertebrate genomes,

effective tools for performing multiple comparisons are required to accurately

and efficiently assist biological discoveries. Here we introduce Mulan

(http://mulan.dcode.org/), a novel method and a network server for comparing

multiple draft and finished-quality sequences to identify functional elements

conserved over evolutionary time. Mulan brings together several novel algorithms:

the TBA multi-aligner program for rapid identification of local sequence

conservation, and the multiTF program for detecting evolutionarily conserved

transcription factor binding sites in multiple alignments. In addition, Mulan

supports two-way communication with the GALA database; alignments of multiple

species dynamically generated in GALA can be viewed in Mulan, and conserved

transcription factor binding sites identified with Mulan/multiTF can be

integrated and overlaid with extensive genome annotation data using GALA. Local

multiple alignments computed by Mulan ensure reliable representation of short-

and large-scale genomic rearrangements in distant organisms. Mulan allows for

interactive modification of critical conservation parameters to differentially

predict conserved regions in comparisons of both closely and distantly related

species. We illustrate the uses and applications of the Mulan tool through

multispecies comparisons of the GATA3 gene locus and the identification of

elements that are conserved in a different way in avians than in other genomes,

allowing speculation on the evolution of birds. Source code for the aligners and

the aligner-evaluation software can be freely downloaded from

http://www.bx.psu.edu/miller\_lab/.

DOI: 10.1101/gr.3007205

PMCID: PMC540288

PMID: 15590941 [Indexed for MEDLINE]

3209. In Silico Biol. 2005;5(4):419-23.

KEGG-based pathway visualization tool for complex omics data.

Arakawa K(1), Kono N, Yamada Y, Mori H, Tomita M.

Author information:

(1)Institute for Advanced Biosciences, Keio University, Endo 5322, Fujisawa,

Kanagawa 252-8520, Japan.

Pathway-level visualization of omics data provides an essential means for systems

biology, to capture the systematic properties of the inner activities of cells.

Here we describe a web-based resource consisting of a web-application for the

visualization of complex omics data onto KEGG pathways to overview all entities

in the context of cellular pathways, and databases created with the software to

visualize a series of microarray data. The web-application accepts transcriptome,

proteome, metabolome, or the combination of these data as input, and because of

this scalability it is advantageous for the visualization of cell simulation

results. The web server can be accessed at

http://www.g-language.org/data/marray/.

PMID: 16268787 [Indexed for MEDLINE]

3210. In Silico Biol. 2005;5(3):341-6. Epub 2005 Apr 11.

In silico simulation of fingerprinting techniques based on double endonuclease

digestion of genomic DNA.

San Millán R(1), Garaizar J, Bikandi J.

Author information:

(1)Department of Immunology, Microbiology and Parasitology, University of the

Basque Country, Paseo de la Universidad, 7, 01006 Vitoria-Gasteiz, Spain.

We have developed an online generic tool for simulation of fingerprinting

techniques based on the double endonuclease digestion of DNA. This tool allows

modelling and modifications of already existing techniques, as well as new

theoretical approaches not yet tried in the lab. It allows the use of any

combination of recognition patterns and discrimination of end types yielded by

restriction with non palindromic recognition sizes. Re-creation of experimental

conditions in silico saves time and reduces laboratory costs. This tool allows

simulation of Amplified Fragment Length Polymorphism (AFLP-PCR), Subtracted

Restriction Fingerprinting (SRF), and additional novel fingerprinting techniques.

Simulation may be performed against custom sequences uploaded to the server, or

against all sequenced bacterial genomes. Different endonuclease types may be

selected from a list, or a recognition sequence may be introduced in the form.

After double digestion of DNA, four fragment types are yielded, and the program

allows their customised selection. Selective nucleotides may be used in the

experiment. Scripts for specific simulation of AFLP-PCR and SRF techniques are

available, and both include a suggestion tool for the selection of endonucleases.

This is the first program available for the simulation of SRF

fingerprinting.AVAILABILITY: This free online tool is available at

http://www.in-silico.com/DDF/.

PMID: 16153186 [Indexed for MEDLINE]

3211. In Silico Biol. 2005;5(2):187-98.

Bioinformatics visualization and integration with open standards: the Bluejay

genomic browser.

Turinsky AL(1), Ah-Seng AC, Gordon PM, Stromer JN, Taschuk ML, Xu EW, Sensen CW.

Author information:

(1)University of Calgary, Faculty of Medicine, Sun Center of Excellence for

Visual Genomics, 3330 Hospital Drive NW, Calgary, AB, Canada, T2N 4N1.

We have created a new Java-based integrated computational environment for the

exploration of genomic data, called Bluejay. The system is capable of using

almost any XML file related to genomic data. Non-XML data sources can be accessed

via a proxy server. Bluejay has several features, which are new to

Bioinformatics, including an unlimited semantic zoom capability, coupled with

Scalable Vector Graphics (SVG) outputs; an implementation of the XLink standard,

which features access to MAGPIE Genecards as well as any BioMOBY service

accessible over the Internet; and the integration of gene chip analysis tools

with the functional assignments. The system can be used as a signed web applet,

Web Start, and a local stand-alone application, with or without connection to the

Internet. It is available free of charge and as open source via

http://bluejay.ucalgary.ca.

PMID: 15972014 [Indexed for MEDLINE]

3212. J Chem Inf Model. 2005 Jan-Feb;45(1):177-82.

ZINC--a free database of commercially available compounds for virtual screening.

Irwin JJ(1), Shoichet BK.

Author information:

(1)Department of Pharmaceutical Chemistry, University of California San

Francisco, Genentech Hall, 600 16th Street, San Francisco, California 94143, USA.

A critical barrier to entry into structure-based virtual screening is the lack of

a suitable, easy to access database of purchasable compounds. We have therefore

prepared a library of 727,842 molecules, each with 3D structure, using catalogs

of compounds from vendors (the size of this library continues to grow). The

molecules have been assigned biologically relevant protonation states and are

annotated with properties such as molecular weight, calculated LogP, and number

of rotatable bonds. Each molecule in the library contains vendor and purchasing

information and is ready for docking using a number of popular docking programs.

Within certain limits, the molecules are prepared in multiple protonation states

and multiple tautomeric forms. In one format, multiple conformations are

available for the molecules. This database is available for free download

(http://zinc.docking.org) in several common file formats including SMILES, mol2,

3D SDF, and DOCK flexibase format. A Web-based query tool incorporating a

molecular drawing interface enables the database to be searched and browsed and

subsets to be created. Users can process their own molecules by uploading them to

a server. Our hope is that this database will bring virtual screening libraries

to a wide community of structural biologists and medicinal chemists.

DOI: 10.1021/ci049714+

PMCID: PMC1360656

PMID: 15667143 [Indexed for MEDLINE]

3213. J Struct Funct Genomics. 2005;6(2-3):135-41.

High-throughput protein production for X-ray crystallography and use of size

exclusion chromatography to validate or refute computational biological unit

predictions.

McMullan D(1), Canaves JM, Quijano K, Abdubek P, Nigoghossian E, Haugen J, Klock

HE, Vincent J, Hale J, Paulsen J, Lesley SA.

Author information:

(1)Joint Center for Structural Genomics, Genomics Institute of the Novartis

Research Foundation, 10675 John Jay Hopkins Drive, San Diego, CA 92121-1127, USA.

The production of large numbers of highly purified proteins for X-ray

crystallography is a significant bottleneck in structural genomics. At the Joint

Center for Structural Genomics (JCSG; http://www.jcsg.org), specific automated

protein expression, purification, and analytical methods are being utilized to

study the proteome of Thermotoga maritima. Anion exchange and size exclusion

chromatography (SEC), intended for the production of highly purified proteins,

have been automated and the procedures are described here in detail. Analytical

SEC has been included as a standard quality control test. A biological unit (BU)

is the macromolecule that has been proven or is presumed to be functional.

Correct assignment of BUs from protein structures can be difficult. BU

predictions obtained via the Protein Quaternary Structure file server (PQS;

http://pqs.ebi.ac.uk/) were compared to SEC data for 16 representative T.

maritima proteins whose structures were solved at the JCSG, revealing an

inconsistency in five cases. Herein, we report that SEC can be used to validate

or disprove PQS-derived oligomeric models. A substantial amount of associated SEC

and structural data should enable us to use certain PQS parameters to gauge the

accuracy of these computational models and to generally improve their

predictions.

DOI: 10.1007/s10969-005-2898-1

PMID: 16211510 [Indexed for MEDLINE]

3214. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D671-4.

VBASE2, an integrative V gene database.

Retter I(1), Althaus HH, Münch R, Müller W.

Author information:

(1)Department of Experimental Immunology, German Research Centre for

Biotechnology, Mascheroder Weg 1, D-38124 Braunschweig, Germany.

The database VBASE2 provides germ-line sequences of human and mouse

immunoglobulin variable (V) genes. It acts as an interconnecting platform between

several existing self-contained data systems: VBASE2 integrates genome sequence

data and links to the V genes in the Ensembl Genome Browser. For a single V gene

sequence, all references to the EMBL nucleotide sequence database are provided,

including references for V(D)J rearrangements. Furthermore, cross-references to

the VBASE database, the IMGT database and the Kabat database are available. A DAS

server allows the display of VBASE2 V genes within the Ensembl Genome Browser.

VBASE2 can be accessed either by a web-based text query or by a sequence

similarity search with the DNAPLOT software. VBASE2 is available at

http://www.vbase2.org, and the DAS server is located at

http://www.dnaplot.com/das.

DOI: 10.1093/nar/gki088

PMCID: PMC540042

PMID: 15608286 [Indexed for MEDLINE]

3215. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D59-66.

HOPPSIGEN: a database of human and mouse processed pseudogenes.

Khelifi A(1), Duret L, Mouchiroud D.

Author information:

(1)Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université

Claude Bernard-Lyon 1, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex,

France. khelifi@biomserv.univ-lyon1.fr

Erratum in

Nucleic Acids Res. 2005;33(1):448. Adel, Khelifi [corrected to Khelifi, Adel];

Laurent, Duret [corrected to Durent, Laurent]; Dominique, Mouchiroud [corrected

to Mouchiroud, Dominique].

Processed pseudogenes result from reverse transcribed mRNAs. In general, because

processed pseudogenes lack promoters, they are no longer functional from the

moment they are inserted into the genome. Subsequently, they freely accumulate

substitutions, insertions and deletions. Moreover, the ancestral structure of

processed pseudogenes could be easily inferred using the sequence of their

functional homologous genes. Owing to these characteristics, processed

pseudogenes represent good neutral markers for studying genome evolution.

Recently, there is an increasing interest for these markers, particularly to help

gene prediction in the field of genome annotation, functional genomics and genome

evolution analysis (patterns of substitution). For these reasons, we have

developed a method to annotate processed pseudogenes in complete genomes. To make

them useful to different fields of research, we stored them in a nucleic acid

database after having annotated them. In this work, we screened both mouse and

human complete genomes from ENSEMBL to find processed pseudogenes generated from

functional genes with introns. We used a conservative method to detect processed

pseudogenes in order to minimize the rate of false positive sequences. Within

processed pseudogenes, some are still having a conserved open reading frame and

some have overlapping gene locations. We designated as retroelements all reverse

transcribed sequences and more strictly, we designated as processed pseudogenes,

all retroelements not falling in the two former categories (having a conserved

open reading or overlapping gene locations). We annotated 5823 retroelements

(5206 processed pseudogenes) in the human genome and 3934 (3428 processed

pseudogenes) in the mouse genome. Compared to previous estimations, the total

number of processed pseudogenes was underestimated but the aim of this procedure

was to generate a high-quality dataset. To facilitate the use of processed

pseudogenes in studying genome structure and evolution, DNA sequences from

processed pseudogenes, and their functional reverse transcribed homologs, are now

stored in a nucleic acid database, HOPPSIGEN. HOPPSIGEN can be browsed on the

PBIL (Pole Bioinformatique Lyonnais) World Wide Web server

(http://pbil.univ-lyon1.fr/) or fully downloaded for local installation.

DOI: 10.1093/nar/gki084

PMCID: PMC540038

PMID: 15608268 [Indexed for MEDLINE]

3216. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D550-2.

5'SAGE: 5'-end Serial Analysis of Gene Expression database.

Kasai Y(1), Hashimoto S, Yamada T, Sese J, Sugano S, Matsushima K, Morishita S.

Author information:

(1)Department of Computational Biology, Graduate School of Frontier Science, The

University of Tokyo, 5-1-5 Kashinoha, Kashiwa City, Chiba 277-8562, Japan.

To comprehensively identify transcription start sites and the frequencies of

individual mRNAs in human cell libraries, a method of 5' end Serial Analysis of

Gene Expression (SAGE) was developed recently, which makes it possible to collect

a large amount of start site information, and subsequently, we have established a

related database server called 5'SAGE. This database displays the observed

frequencies of individual 5' end SAGE tags and previously unknown transcription

start sites in the promoter regions, introns and intergenic regions of known

genes. 5'SAGE will be useful for analyzing promoter regions and start site

variation in different tissues, and is freely available at

http://5sage.gi.k.u-tokyo.ac.jp/.

DOI: 10.1093/nar/gki085

PMCID: PMC540039

PMID: 15608259 [Indexed for MEDLINE]

3217. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D523-6.

IPD--the Immuno Polymorphism Database.

Robinson J(1), Waller MJ, Stoehr P, Marsh SG.

Author information:

(1)Anthony Nolan Research Institute, Royal Free Hospital, Pond Street, Hampstead,

London NW3 2QG, UK.

The Immuno Polymorphism Database (IPD) (http://www.ebi.ac.uk/ipd/) is a set of

specialist databases related to the study of polymorphic genes in the immune

system. IPD currently consists of four databases: IPD-KIR, contains the allelic

sequences of Killer-cell Immunoglobulin-like Receptors; IPD-MHC, a database of

sequences of the Major Histocompatibility Complex of different species; IPD-HPA,

alloantigens expressed only on platelets; and IPD-ESTAB, which provides access to

the European Searchable Tumour Cell-Line Database, a cell bank of immunologically

characterized melanoma cell lines. The IPD project works with specialist groups

or nomenclature committees who provide and curate individual sections before they

are submitted to IPD for online publication. The IPD project stores all the data

in a set of related databases. Those sections with similar data, such as IPD-KIR

and IPD-MHC share the same database structure. The sharing of a common database

structure makes it easier to implement common tools for data submission and

retrieval. The data are currently available online from the website and ftp

directory; files will also be made available in different formats to download

from the website and ftp server. The data will also be included in SRS, BLAST and

FASTA search engines at the European Bioinformatics Institute.

DOI: 10.1093/nar/gki032

PMCID: PMC539986

PMID: 15608253 [Indexed for MEDLINE]

3218. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D321-4.

Metagrowth: a new resource for the building of metabolic hypotheses in

microbiology.

Ogata H(1), Claverie JM.

Author information:

(1)Information Génomique et Structurale, CNRS UPR 2589, 31 Chemin Joseph Aiguier,

13402 Marseille Cedex 20, France. Hiroyuki.Ogata@igs.cnrs-mrs.fr

Metagrowth is a new type of knowledge base developed to guide the experimental

studies of culture conditions of obligate parasitic bacteria. We have gathered

biological evidences giving possible clues to the development of the axenic (i.e.

'cell-free') growth of obligate parasites from various sources including

published literature, genomic sequence information, metabolic databases and

transporter databases. The database entries are composed of those evidences and

specific hypotheses derived from them. Currently, 200 entries are available for

Rickettsia prowazekii, Rickettsia conorii, Tropheryma whipplei, Treponema

pallidum, Mycobacterium tuberculosis and Coxiella burnetii. The web interface of

Metagrowth helps users to design new axenic culture media eventually suitable for

those bacteria. Metagrowth is accessible at

http://igs-server.cnrs-mrs.fr/axenic/.

DOI: 10.1093/nar/gki042

PMCID: PMC539996

PMID: 15608207 [Indexed for MEDLINE]

3219. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D289-93.

Biomedical term mapping databases.

Wren JD(1), Chang JT, Pustejovsky J, Adar E, Garner HR, Altman RB.

Author information:

(1)Advanced Center for Genome Technology, Department of Botany and Microbiology,

The University of Oklahoma, 101 David L. Boren Blvd, Rm 2025, Norman, OK 73019,

USA. Jonathan.Wren@OU.edu

Longer words and phrases are frequently mapped onto a shorter form such as

abbreviations or acronyms for efficiency of communication. These abbreviations

are pervasive in all aspects of biology and medicine and as the amount of

biomedical literature grows, so does the number of abbreviations and the average

number of definitions per abbreviation. Even more confusing, different authors

will often abbreviate the same word/phrase differently. This ambiguity impedes

our ability to retrieve information, integrate databases and mine textual

databases for content. Efforts to standardize nomenclature, especially those

doing so retrospectively, need to be aware of different abbreviatory mappings and

spelling variations. To address this problem, there have been several efforts to

develop computer algorithms to identify the mapping of terms between short and

long form within a large body of literature. To date, four such algorithms have

been applied to create online databases that comprehensively map biomedical terms

and abbreviations within MEDLINE: ARGH (http://lethargy.swmed.edu/ARGH/argh.asp),

the Stanford Biomedical Abbreviation Server

(http://bionlp.stanford.edu/abbreviation/), AcroMed

(http://medstract.med.tufts.edu/acro1.1/index.htm) and SaRAD

(http://www.hpl.hp.com/research/idl/projects/abbrev.html). In addition to serving

as useful computational tools, these databases serve as valuable references that

help biologists keep up with an ever-expanding vocabulary of terms.

DOI: 10.1093/nar/gki137

PMCID: PMC540091

PMID: 15608198 [Indexed for MEDLINE]

3220. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D247-51.

The CATH Domain Structure Database and related resources Gene3D and DHS provide

comprehensive domain family information for genome analysis.

Pearl F(1), Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden

R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I,

Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C.

Author information:

(1)Biochemistry and Molecular Biology Department, University College London,

University of London, Gower Street, London WC1E 6BT, UK.

The CATH database of protein domain structures

(http://www.biochem.ucl.ac.uk/bsm/cath/) currently contains 43,229 domains

classified into 1467 superfamilies and 5107 sequence families. Each structural

family is expanded with sequence relatives from GenBank and completed genomes,

using a variety of efficient sequence search protocols and reliable thresholds.

This extended CATH protein family database contains 616,470 domain sequences

classified into 23,876 sequence families. This results in the significant

expansion of the CATH HMM model library to include models built from the CATH

sequence relatives, giving a 10% increase in coverage for detecting remote

homologues. An improved Dictionary of Homologous superfamilies (DHS)

(http://www.biochem.ucl.ac.uk/bsm/dhs/) containing specific sequence, structural

and functional information for each superfamily in CATH considerably assists

manual validation of homologues. Information on sequence relatives in CATH

superfamilies, GenBank and completed genomes is presented in the CATH associated

DHS and Gene3D resources. Domain partnership information can be obtained from

Gene3D (http://www.biochem.ucl.ac.uk/bsm/cath/Gene3D/). A new CATH server has

been implemented (http://www.biochem.ucl.ac.uk/cgi-bin/cath/CathServer.pl)

providing automatic classification of newly determined sequences and structures

using a suite of rapid sequence and structure comparison methods. The statistical

significance of matches is assessed and links are provided to the putative

superfamily or fold group to which the query sequence or structure is assigned.

DOI: 10.1093/nar/gki024

PMCID: PMC539978

PMID: 15608188 [Indexed for MEDLINE]

3221. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D226-9.

The SYSTERS Protein Family Database in 2005.

Meinel T(1), Krause A, Luz H, Vingron M, Staub E.

Author information:

(1)Computational Molecular Biology Department, Max Planck Institute for Molecular

Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany.

The SYSTERS project aims to provide a meaningful partitioning of the whole

protein sequence space by a fully automatic procedure. A refined two-step

algorithm assigns each protein to a family and a superfamily. The sequence data

underlying SYSTERS release 4 now comprise several protein sequence databases

derived from completely sequenced genomes (ENSEMBL, TAIR, SGD and GeneDB), in

addition to the comprehensive Swiss-Prot/TrEMBL databases. The SYSTERS web server

(http://systers.molgen.mpg.de) provides access to 158 153 SYSTERS protein

families. To augment the automatically derived results, information from external

databases like Pfam and Gene Ontology are added to the web server. Furthermore,

users can retrieve pre-processed analyses of families like multiple alignments

and phylogenetic trees. New query options comprise a batch retrieval tool for

functional inference about families based on automatic keyword extraction from

sequence annotations. A new access point, PhyloMatrix, allows the retrieval of

phylogenetic profiles of SYSTERS families across organisms with completely

sequenced genomes.

DOI: 10.1093/nar/gki030

PMCID: PMC539984

PMID: 15608183 [Indexed for MEDLINE]

3222. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D212-5.

The ProDom database of protein domain families: more emphasis on 3D.

Bru C(1), Courcelle E, Carrère S, Beausse Y, Dalmar S, Kahn D.

Author information:

(1)Laboratoire des Interactions Plantes-Microorganismes, INRA/CNRS, BP 27,

F-31326 Castanet-Tolosan Cedex, France.

ProDom is a comprehensive database of protein domain families generated from the

global comparison of all available protein sequences. Recent improvements include

the use of three-dimensional (3D) information from the SCOP database; a

completely redesigned web interface (http://www.toulouse.inra.fr/prodom.html);

visualization of ProDom domains on 3D structures; coupling of ProDom analysis

with the Geno3D homology modelling server; Bayesian inference of evolutionary

scenarios for ProDom families. In addition, we have developed ProDom-SG, a

ProDom-based server dedicated to the selection of candidate proteins for

structural genomics.

DOI: 10.1093/nar/gki034

PMCID: PMC539988

PMID: 15608179 [Indexed for MEDLINE]

3223. Proteins. 2005;61 Suppl 7:157-66.

Prediction of CASP6 structures using automated Robetta protocols.

Chivian D(1), Kim DE, Malmström L, Schonbrun J, Rohl CA, Baker D.

Author information:

(1)Department of Biochemistry, University of Washington, Seattle, Washington

98195, USA.

The Robetta server and revised automatic protocols were used to predict

structures for CASP6 targets. Robetta is a publicly available protein structure

prediction server (http://robetta.bakerlab.org/ that uses the Rosetta de novo and

homology modeling structure prediction methods. We incorporated some of the

lessons learned in the CASP5 experiment into the server prior to participating in

CASP6. We additionally tested new ideas that were amenable to full-automation

with an eye toward improving the server. We find that the Robetta server shows

the greatest promise for the more challenging targets. The most significant

finding from CASP5, that automated protocols can be roughly comparable in ability

with the better human-intervention predictors, is repeated here in CASP6.

2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20733

PMID: 16187358 [Indexed for MEDLINE]

3224. Proteins. 2005;61 Suppl 7:152-6.

SPARKS 2 and SP3 servers in CASP6.

Zhou H(1), Zhou Y.

Author information:

(1)Howard Hughes Medical Institute Center for Single Molecule Biophysics,

Department of Physiology and Biophysics, State University of New York, Buffalo,

New York 14214, USA.

Two single-method servers, SPARKS 2 and SP3, participated in automatic-server

predictions in CASP6. The overall results for all as well as detailed performance

in comparative modeling targets are presented. It is shown that both SPARKS 2 and

SP3 are able to recognize their corresponding best templates for all easy

comparative modeling targets. The alignment accuracy, however, is not always the

best among all the servers. Possible factors are discussed. SPARKS 2 and SP3 fold

recognition servers, as well as their executables, are freely available for all

academic users on http://theory.med.buffalo.edu.

2005 Wiley-Liss, Inc.

DOI: 10.1002/prot.20732

PMID: 16187357 [Indexed for MEDLINE]

3225. Proteins. 2005 Jan 1;58(1):190-9.

MSDsite: a database search and retrieval system for the analysis and viewing of

bound ligands and active sites.

Golovin A(1), Dimitropoulos D, Oldfield T, Rachedi A, Henrick K.

Author information:

(1)EMBL Outstation, The European Bioinformatics Institute, Welcome Trust Genome

Campus, Hinxton, Cambridge, United Kingdom.

The three-dimensional environments of ligand binding sites have been derived from

the parsing and loading of the PDB entries into a relational database. For each

bound molecule the biological assembly of the quaternary structure has been used

to determine all contact residues and a fast interactive search and retrieval

system has been developed. Prosite pattern and short sequence search options are

available together with a novel graphical query generator for inter-residue

contacts. The database and its query interface are accessible from the Internet

through a web server located at: http://www.ebi.ac.uk/msd-srv/msdsite.

DOI: 10.1002/prot.20288

PMID: 15468317 [Indexed for MEDLINE]

3226. Stud Health Technol Inform. 2005;116:290-5.

Trends in free WWW-based E-learning Modules seen from the Learning Resource

Server Medicine (LRSMed).

Stausberg J(1), Geueke M, Bludßat K.

Author information:

(1)Institute for Medical Informatics, Biometry and Epidemiology, Medical Faculty,

University of Duisburg-Essen, Germany. stausberg@uni-essen.de

Despite the lost enthusiasm concerning E-learning a lot of material is available

on the World Wide Web (WWW) free of charge. This material is collected and

systematically described by services like the Learning Resource Server Medicine

(LRSMed) at http://mmedia.medizin.uni-essen.de/portal/. With the LRSMed

E-learning modules are made available for medical students by means of a metadata

description that can be used for a catalogue search. The number of resources

included has risen enormously from 100 in 1999 up to 805 today. Especially in

2004 there was an exponential increase in the LRSMed's content. Anatomy is still

the field with the highest amount of available material, but general medicine has

improved its position over the years and is now the second one. Technically and

didactically simple material as scripts, textbooks, and link lists (called info

services) is still dominating. Similar to 1999, there is not one module which

could be truly referred to as tutorial dialogue. Simple material can not replace

face-to-face-teaching. But it could be combined with conventional courses to

establish some kind of blending learning. The scene of free E-learning modules on

the WWW is ready to meet current challenges for efficient training of students

and continuing education in medicine.

PMID: 16160274 [Indexed for MEDLINE]

3227. Toxicol Pathol. 2005;33(5):517-32.

Chemical-induced atrial thrombosis in NTP rodent studies.

Yoshizawa K(1), Kissling GE, Johnson JA, Clayton NP, Flagler ND, Nyska A.

Author information:

(1)Laboratory of Experimental Pathology, National Institute of Environmental

Health Sciences (NIEHS), Research Triangle Park, North Carolina 27709, USA.

Cardiac thrombosis, one of the causes of sudden death throughout the world, plays

a principal role in several cardiovascular diseases, such as myocardial

infarction and stroke in humans. Data from studies of induction of chemical

thrombosis in rodents help to identify substances in our environment that may

contribute to cardiac thrombosis. Results for more than 500 chemicals tested in

rodents in 2-year bioassays have been published as Technical Reports of the

National Toxicology Program (NTP) http://ntp-server.niehs.nih.gov/index. We

evaluated atrial thrombosis induced by these chemical exposures and compared it

to similarly induced lesions reported in the literature. Spontaneous rates of

cardiac thrombosis were determined for control Fischer 344 rats and B6C3F1 mice:

0% in rats and mice in 90-day studies and, in 2-year studies, 0.7% in both

genders of mice, 4% in male rats, and 1% in female rats. Incidences of atrial

thrombosis were increased in high-dosed groups involving 13 compounds (incidence

rate: 20-100%): 2-butoxyethanol, C.I. Direct Blue 15, bis(2-chloroethoxy)methane,

diazoaminobenzene, diethanolamine, 3,3'-dimethoxybenzidine dihydrochloride,

hexachloroethane, isobutene, methyleugenol, oxazepam, C.I. Pigment Red 23, C.I.

Acid Red 114, and 4,4'-thiobis(6-t-butyl-m-cresol). The main localization of

spontaneously occurring and chemically induced thromboses occurred in the left

atrium. The literature survey suggested that chemical-induced atrial thrombosis

might be closely related to myocardial injury, endothelial injury, circulatory

stasis, hypercoagulability, and impaired atrial mechanical activity, such as

atrial fibrillation, which could cause stasis of blood within the left atrial

appendage, contributing to left atrial thrombosis. Supplementary data referenced

in this paper are not printed in this issue of Toxicologic Pathology. They are

available as downloadable files at

http://taylorandfrancis.metapress.com/openurl.asp?genre=journal&issn=0192-6233.

To access them, click on the issue link for 33(5), then select this article. A

download option appears at the bottom of this abstract. In order to access the

full article online, you must either have an individual subscription or a member

subscription accessed through www.toxpath.org.

DOI: 10.1080/01926230591034429

PMID: 16048847 [Indexed for MEDLINE]

3228. Bioinformatics. 2004 Dec 12;20(18):3662-4. Epub 2004 Jul 15.

Interactive gene-order comparison for multiple small genomes.

Kaluszka A(1), Gibas C.

Author information:

(1)Department of Biology, College of Science, Virginia Polytechnic Institute and

State University, Blacksburg, VA 24061, USA.

The Genome Organization Analysis Tool (GOAT) is a program that performs

comparative sequence analysis on ordered gene lists from annotated genomes,

provides visual and tabular output, and provides means of accessing and analyzing

related gene sequence data, for the purpose of comparing genome organization at

the gene-order level. GOAT can be used to compare any two or more genomes or

chromosomes on demand, or configured to provide access to precomputed comparisons

of a specific group of genome sequences.AVAILABILITY: Demonstration web server

and software download, subject to the Virginia Tech Noncommercial License are

available at http://gaia.biotech.vt.edu/goat/.

SUPPLEMENTARY INFORMATION: Updates, installation and configuration information

are available at http://gaia.biotech.vt.edu/goat.

DOI: 10.1093/bioinformatics/bth406

PMID: 15256414 [Indexed for MEDLINE]

3229. Bioinformatics. 2004 Dec 12;20(18):3665-7. Epub 2004 Jul 15.

CRH\_Server: an online comparative and radiation hybrid mapping server for the

canine genome.

Hitte C(1), Derrien T, André C, Ostrander EA, Galibert F.

Author information:

(1)UMR 6061 CNRS, Génétique et Développement, faculté de Médecine, Université de

Rennes1, 2 avenue Pr. Léon Bernard, 35043 Rennes Cédex, France.

hitte@univ-rennes1.fr <hitte@univ-rennes1.fr>

CRH\_Server is an on line Comparative and Radiation Hybrid mapping Server

dedicated to canine genomics. CRH\_Server allows users to compute their own RH

data using the current canine RH map, and allows comparative dog/human mapping

analyses. Finally, it suggests multiple options for storage and queries of the

dog RH database.AVAILABILITY:

http://idefix.univ-rennes1.fr:8080/Dogs/rh-server.html.

SUPPLEMENTARY INFORMATION: All information is available at

http://idefix.univ-rennes1.fr:8080/Dogs/help\_rh-server.html.

DOI: 10.1093/bioinformatics/bth411

PMID: 15256409 [Indexed for MEDLINE]

3230. Bioinformatics. 2004 Dec 12;20(18):3682-6. Epub 2004 Jul 15.

CBS Genome Atlas Database: a dynamic storage for bioinformatic results and

sequence data.

Hallin PF(1), Ussery DW.

Author information:

(1)Center for Biological Sequence Analysis (CBS), BioCentrum-DTU, Building 208,

The Technical University of Denmark, DK-2800 Lyngby, Denmark. pfh@cbs.dtu.dk

<pfh@cbs.dtu.dk>

Currently, new bacterial genomes are being published on a monthly basis. With the

growing amount of genome sequence data, there is a demand for a flexible and

easy-to-maintain structure for storing sequence data and results from

bioinformatic analysis. More than 150 sequenced bacterial genomes are now

available, and comparisons of properties for taxonomically similar organisms are

not readily available to many biologists. In addition to the most basic

information, such as AT content, chromosome length, tRNA count and rRNA count, a

large number of more complex calculations are needed to perform detailed

comparative genomics. DNA structural calculations like curvature and stacking

energy, DNA compositions like base skews, oligo skews and repeats at the local

and global level are just a few of the analysis that are presented on the CBS

Genome Atlas Web page. Complex analysis, changing methods and frequent addition

of new models are factors that require a dynamic database layout. Using basic

tools like the GNU Make system, csh, Perl and MySQL, we have created a flexible

database environment for storing and maintaining such results for a collection of

complete microbial genomes. Currently, these results counts to more than 220

pieces of information. The backbone of this solution consists of a program

package written in Perl, which enables administrators to synchronize and update

the database content. The MySQL database has been connected to the CBS web-server

via PHP4, to present a dynamic web content for users outside the center. This

solution is tightly fitted to existing server infrastructure and the solutions

proposed here can perhaps serve as a template for other research groups to solve

database issues.AVAILABILITY: A web based user interface which is dynamically

linked to the Genome Atlas Database can be accessed via

www.cbs.dtu.dk/services/GenomeAtlas/.

SUPPLEMENTARY INFORMATION: This paper has a supplemental information page which

links to the examples presented:

www.cbs.dtu.dk/services/GenomeAtlas/suppl/bioinfdatabase.

DOI: 10.1093/bioinformatics/bth423

PMID: 15256401 [Indexed for MEDLINE]

3231. Bioinformatics. 2004 Dec 12;20(18):3647-51. Epub 2004 Jul 9.

CSB.DB: a comprehensive systems-biology database.

Steinhauser D(1), Usadel B, Luedemann A, Thimm O, Kopka J.

Author information:

(1)Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476

Golm, Germany. Steinhauser@mpimp-golm.mpg.de <Steinhauser@mpimp-golm.mpg.de>

SUMMARY: The open access comprehensive systems-biology database (CSB.DB) presents

the results of bio-statistical analyses on gene expression data in association

with additional biochemical and physiological knowledge. The main aim of this

database platform is to provide tools that support insight into life's complexity

pyramid with a special focus on the integration of data from transcript and

metabolite profiling experiments. The central part of CSB.DB, which we describe

in this applications note, is a set of co-response databases that currently focus

on the three key model organisms, Escherichia coli, Saccharomyces cerevisiae and

Arabidopsis thaliana. CSB.DB gives easy access to the results of large-scale

co-response analyses, which are currently based exclusively on the publicly

available compendia of transcript profiles. By scanning for the best co-responses

among changing transcript levels, CSB.DB allows to infer hypotheses on the

functional interaction of genes. These hypotheses are novel and not accessible

through analysis of sequence homology. The database enables the search for pairs

of genes and larger units of genes, which are under common transcriptional

control. In addition, statistical tools are offered to the user, which allow

validation and comparison of those co-responses that were discovered by gene

queries performed on the currently available set of pre-selectable datasets.

AVAILABILITY: All co-response databases can be accessed through the CSB.DB Web

server (http://csbdb.mpimp-golm.mpg.de/).

DOI: 10.1093/bioinformatics/bth398

PMID: 15247097 [Indexed for MEDLINE]

3232. Bioinformatics. 2004 Dec 12;20(18):3656-8. Epub 2004 Jul 9.

DNMAD: web-based diagnosis and normalization for microarray data.

Vaquerizas JM(1), Dopazo J, Díaz-Uriarte R.

Author information:

(1)Bioinformatics Unit, CNIO, Spanish National Cancer Centre, Melchor Fernández

Almagro 3, 28029 Madrid, Spain.

SUMMARY: We present a web server for Diagnosis and Normalization of MicroArray

Data (DNMAD). DNMAD includes several common data transformations such as spatial

and global robust local regression or multiple slide normalization, and allows

for detecting several kinds of errors that result from the manipulation and the

image analysis of the arrays. This tool offers a user-friendly interface, and is

completely integrated within the Gene Expression Pattern Analysis Suite (GEPAS).

AVAILABILITY: The tool is accessible on-line at http://dnmad.bioinfo.cnio.es.

DOI: 10.1093/bioinformatics/bth401

PMID: 15247094 [Indexed for MEDLINE]

3233. BMC Bioinformatics. 2004 Dec 10;5:195.

GECKO: a complete large-scale gene expression analysis platform.

Theilhaber J(1), Ulyanov A, Malanthara A, Cole J, Xu D, Nahf R, Heuer M, Brockel

C, Bushnell S.

Author information:

(1)Cambridge Genomics Center, Sanofi-Aventis, 26 Landsdowne Street, Cambridge, MA

02139, USA. joachim.theilhaber@aventis.com

BACKGROUND: Gecko (Gene Expression: Computation and Knowledge Organization) is a

complete, high-capacity centralized gene expression analysis system, developed in

response to the needs of a distributed user community.

RESULTS: Based on a client-server architecture, with a centralized repository of

typically many tens of thousands of Affymetrix scans, Gecko includes automatic

processing pipelines for uploading data from remote sites, a data base, a

computational engine implementing approximately 50 different analysis tools, and

a client application. Among available analysis tools are clustering methods,

principal component analysis, supervised classification including feature

selection and cross-validation, multi-factorial ANOVA, statistical contrast

calculations, and various post-processing tools for extracting data at given

error rates or significance levels. On account of its open architecture, Gecko

also allows for the integration of new algorithms. The Gecko framework is very

general: non-Affymetrix and non-gene expression data can be analyzed as well. A

unique feature of the Gecko architecture is the concept of the Analysis Tree

(actually, a directed acyclic graph), in which all successive results in ongoing

analyses are saved. This approach has proven invaluable in allowing a large

(approximately 100 users) and distributed community to share results, and to

repeatedly return over a span of years to older and potentially very complex

analyses of gene expression data.

CONCLUSIONS: The Gecko system is being made publicly available as free software

http://sourceforge.net/projects/geckoe. In totality or in parts, the Gecko

framework should prove useful to users and system developers with a broad range

of analysis needs.

DOI: 10.1186/1471-2105-5-195

PMCID: PMC539353

PMID: 15588317 [Indexed for MEDLINE]

3234. Acta Crystallogr D Biol Crystallogr. 2004 Dec;60(Pt 12 Pt 1):2240-9. Epub 2004

Nov 26.

The Uppsala Electron-Density Server.

Kleywegt GJ(1), Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA.

Author information:

(1)Department of Cell and Molecular Biology, Uppsala University, Biomedical

Centre, Box 596, SE-751 24 Uppsala, Sweden. gerard@xray.bmc.uu.se

The Uppsala Electron Density Server (EDS; http://eds.bmc.uu.se/) is a web-based

facility that provides access to electron-density maps and statistics concerning

the fit of crystal structures and their maps. Maps are available for

approximately 87% of the crystallographic Protein Data Bank (PDB) entries for

which structure factors have been deposited and for which straightforward map

calculations succeed in reproducing the published R value to within five

percentage points. Here, an account is provided of the methods that are used to

generate the information contained in the server. Some of the problems that are

encountered in the map-generation process as well as some spin-offs of the

project are also discussed.

DOI: 10.1107/S0907444904013253

PMID: 15572777 [Indexed for MEDLINE]

3235. Comb Chem High Throughput Screen. 2004 Dec;7(8):757-61.

Ligand.Info small-molecule Meta-Database.

von Grotthuss M(1), Koczyk G, Pas J, Wyrwicz LS, Rychlewski L.

Author information:

(1)BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznan, Poland.

Ligand.Info is a compilation of various publicly available databases of small

molecules. The total size of the Meta-Database is over 1 million entries. The

compound records contain calculated three-dimensional coordinates and sometimes

information about biological activity. Some molecules have information about FDA

drug approving status or about anti-HIV activity. Meta-Database can be downloaded

from the http://Ligand.Info web page. The database can also be screened using a

Java-based tool. The tool can interactively cluster sets of molecules on the user

side and automatically download similar molecules from the server. The

application requires the Java Runtime Environment 1.4 or higher, which can be

automatically downloaded from Sun Microsystems or Apple Computer and installed

during the first use of Ligand.Info on desktop systems, which support Java (Ms

Windows, Mac OS, Solaris, and Linux). The Ligand.Info Meta-Database can be used

for virtual high-throughput screening of new potential drugs. Presented examples

showed that using a known antiviral drug as query the system was able to find

others antiviral drugs and inhibitors.

PMID: 15578937 [Indexed for MEDLINE]

3236. Comput Biol Chem. 2004 Dec;28(5-6):367-74.

Comparing two K-category assignments by a K-category correlation coefficient.

Gorodkin J(1).

Author information:

(1)Center for Bioinformatics and Division of Genetics, IBHV, The Royal Veterinary

and Agricultural University, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark.

Predicted assignments of biological sequences are often evaluated by Matthews

correlation coefficient. However, Matthews correlation coefficient applies only

to cases where the assignments belong to two categories, and cases with more than

two categories are often artificially forced into two categories by considering

what belongs and what does not belong to one of the categories, leading to the

loss of information. Here, an extended correlation coefficient that applies to

K-categories is proposed, and this measure is shown to be highly applicable for

evaluating prediction of RNA secondary structure in cases where some predicted

pairs go into the category "unknown" due to lack of reliability in predicted

pairs or unpaired residues. Hence, predicting base pairs of RNA secondary

structure can be a three-category problem. The measure is further shown to be

well in agreement with existing performance measures used for ranking protein

secondary structure predictions. Server and software is available at

http://rk.kvl.dk/.

DOI: 10.1016/j.compbiolchem.2004.09.006

PMID: 15556477 [Indexed for MEDLINE]

3237. J Clin Pathol. 2004 Dec;57(12):1288-91.

A digital atlas of breast histopathology: an application of web based virtual

microscopy.

Lundin M(1), Lundin J, Helin H, Isola J.

Author information:

(1)Institute of Medical Technology, University and University Hospital of

Tampere, 33014 Tampere, Finland.

Comment in

J Clin Pathol. 2004 Dec;57(12):1250-1.

AIMS: To develop an educationally useful atlas of breast histopathology, using

advanced web based virtual microscopy technology.

METHODS: By using a robotic microscope and software adopted and modified from the

aerial and satellite imaging industry, a virtual microscopy system was developed

that allows fully automated slide scanning and image distribution via the

internet. More than 150 slides were scanned at high resolution with an oil

immersion x 40 objective (numerical aperture, 1.3) and archived on an image

server residing in a high speed university network.

RESULTS: A publicly available website was constructed,

http://www.webmicroscope.net/breastatlas, which features a comprehensive virtual

slide atlas of breast histopathology according to the World Health Organisation

2003 classification. Users can view any part of an entire specimen at any

magnification within a standard web browser. The virtual slides are supplemented

with concise textual descriptions, but can also be viewed without diagnostic

information for self assessment of histopathology skills.

CONCLUSIONS: Using the technology described here, it is feasible to develop

clinically and educationally useful virtual microscopy applications. Web based

virtual microscopy will probably become widely used at all levels in pathology

teaching.

DOI: 10.1136/jcp.2004.018739

PMCID: PMC1770524

PMID: 15563669 [Indexed for MEDLINE]

3238. Proteomics. 2004 Dec;4(12):3864-80.

Activation and expression of proteins during synchronous germination of aerial

spores of Streptomyces granaticolor.

Bobek J(1), Halada P, Angelis J, Vohradský J, Mikulík K.

Author information:

(1)Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech

Republic.

Synchronously germinating aerial spores of Streptomyces granaticolor were used to

study protein activation and expression during the transition from dormant to

metabolically active vegetative forms. The first phase of protein activation is

associated with the solubility of proteins. Three major chaperones, DnaK, Trigger

factor, and GroEL, were identified in spores. Enhancement in rate of protein

synthesis during germination was accompanied by the association of TF and DnaK

with ribosomes. During germination, the chaperones TF, GroEL, and DnaK undergo

reversible phosphorylation. GroEL was phosphorylated on both Ser and Thr, whereas

phosphorylation of DnaK and TF was detected on Thr only. A proteomic approach was

used to gain more information on protein expression during germination on two

types of media differing in the ability of cells to produce antibiotic

granaticin. To obtain an overview of the metabolic activity of germinating

spores, glycolytic enzymes, enzymes of citric acid cycle, metabolism of amino

acids and nucleic acids, and components of the protein synthesis system were

identified and analyzed using the proteomic database. The results were deposited

on the SWICZ proteomic server and are accessible on http://proteom.biomed.cas.cz.

DOI: 10.1002/pmic.200400818

PMID: 15378695 [Indexed for MEDLINE]

3239. Bioinformatics. 2004 Nov 1;20(16):2751-8. Epub 2004 May 14.

A neural network method for prediction of beta-turn types in proteins using

evolutionary information.

Kaur H(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector-39A, Chandigarh, India.

MOTIVATION: The prediction of beta-turns is an important element of protein

secondary structure prediction. Recently, a highly accurate neural network based

method Betatpred2 has been developed for predicting beta-turns in proteins using

position-specific scoring matrices (PSSM) generated by PSI-BLAST and secondary

structure information predicted by PSIPRED. However, the major limitation of

Betatpred2 is that it predicts only beta-turn and non-beta-turn residues and does

not provide any information of different beta-turn types. Thus, there is a need

to predict beta-turn types using an approach based on multiple sequence

alignment, which will be useful in overall tertiary structure prediction.

RESULTS: In the present work, a method has been developed for the prediction of

beta-turn types I, II, IV and VIII. For each turn type, two consecutive

feed-forward back-propagation networks with a single hidden layer have been used

where the first sequence-to-structure network has been trained on single

sequences as well as on PSI-BLAST PSSM. The output from the first network along

with PSIPRED predicted secondary structure has been used as input for the

second-level structure-to-structure network. The networks have been trained and

tested on a non-homologous dataset of 426 proteins chains by 7-fold

cross-validation. It has been observed that the prediction performance for each

turn type is improved significantly by using multiple sequence alignment. The

performance has been further improved by using a second level

structure-to-structure network and PSIPRED predicted secondary structure

information. It has been observed that Type I and II beta-turns have better

prediction performance than Type IV and VIII beta-turns. The final network yields

an overall accuracy of 74.5, 93.5, 67.9 and 96.5% with MCC values of 0.29, 0.29,

0.23 and 0.02 for Type I, II, IV and VIII beta-turns, respectively, and is better

than random prediction.

AVAILABILITY: A web server for prediction of beta-turn types I, II, IV and VIII

based on above approach is available at

http://www.imtech.res.in/raghava/betaturns/ and

http://bioinformatics.uams.edu/mirror/betaturns/ (mirror site).

DOI: 10.1093/bioinformatics/bth322

PMID: 15145798 [Indexed for MEDLINE]

3240. Genomics Proteomics Bioinformatics. 2004 Nov;2(4):253-5.

A novel method for N-terminal acetylation prediction.

Liu Y(1), Lin Y.

Author information:

(1)School of Software, Tsinghua University, Beijing 100084, China.

yingliu03@mails.tsinghua.edu.cn

The NetAcet method has been developed to make predictions of N-terminal

acetylation sites, but more information of the data set could be utilized to

improve the performance of the model. By employing a new way to extract patterns

from sequences and using a sample balancing mechanism, we obtained a correlation

coefficient of 0.85, and a sensitivity of 93% on an independent mammalian data

set. A web server utilizing this method has been constructed and is available at

http://166.111.24.5/acetylation.html.

PMCID: PMC5187419

PMID: 15901254 [Indexed for MEDLINE]

3241. Ginekol Pol. 2004 Nov;75(11):896-903.

[The use of interactive medical data transfer to increase effectiveness of a mass

screening program].

[Article in Polish]

Knapp P(1), Zbroch T, Knapp P.

Author information:

(1)Klinika Ginckologii AM, Białymstoku. knapp@pb.bialystok.pl

Cervical cancer is still called the most frequent neoplasm refer to women's

genital tract (incidence rate 30-35/100,000/year). Scandinavian countries

experiences have shown that effective screening program especially used

multimedia (radio, TV, Internet) is one of the based elements, which should

improve situation in prevention of cervical cancer. The aim of this study was an

attempt to create medical Website including specialized and basic informations

for doctors and patients about cervical and breast cancer, their prophylactic and

treatment. An influence of these data to increase active participation in

screening program was also analysed. Material covers a random sample of 1200

women (30-59 years of age) who were invited to participate in cytology screening

in two edition of the program (1997-99, 2000-02). In the years of 2000-02

information about Website was added to the invitation. Medical Website was

created as server information. A special questionnaire prepared by Sociology

Department was distributed to 1200 women. In both group participation during two

screening edition was compared and analysed. All data were statistically

correlated. Website address was created: http://pkzr.ac.bialystok.pl. From the

group of 1200 questioned (mean age--41.8) answered 1059 women (88.25%). 6.98%

(n=74) have never used this Website but 20 of them have done cytology test after

receiving an invitation which were posted to all 1200 women. The last of 985

women analysed Website. 80.4% (n=792) have received enough informations about

cervical and breast cancer (4.67 point) and decided to participate in cervical

cancer screening. From this group only 428 women actively took part in cytology

exams in 1997-99. 189 women participated in the screening programme from the last

193 patients (19.6%) who analysed server informations (4.89 point). From 141

people who have not answered for the questionnaire 23 have done cytology test.

Participation in the second edition of cervical cancer screening programme

(1997-99) was 41.9%. It was really two times lower than in the same group who

could used Internet and was estimated on 85.3%. Internet

site--http://pkzr.ac.bialystok.pl seems to be a basic digital textbook about

cervical and breast neoplastic disease and their prophylactic. It is also

interactive transferring data source that could create healthy attitude to

increase effectiveness of screening programmes.

PMID: 15754581 [Indexed for MEDLINE]

3242. J Zhejiang Univ Sci. 2004 Nov;5(11):1367-73.

A multi-agent system architecture for geographic information gathering.

Gao GY(1), Wang SK.

Author information:

(1)College of Computer Science, Zhejiang University, Hangzhou 310027, China.

ggy0222@sina.com.

World Wide Web (WWW) is a vast repository of information, including a great deal

of geographic information. But the location and retrieval of geographic

information will require a significant amount of time and effort. In addition,

different users usually have different views and interests in the same

information. To resolve such problems, this paper first proposed a model of

geographic information gathering based on multi-Agent (MA) architecture. Then

based on this model, we construct a prototype system with GML (Geography Markup

Language). This system consists of three tiers-Client, Web Server and Data

Resource. Finally, we expatiate on the process of Web Server.

DOI: 10.1631/jzus.2004.1367

PMID: 15495329 [Indexed for MEDLINE]

3243. BMC Bioinformatics. 2004 Oct 28;5:167.

Improvement of alignment accuracy utilizing sequentially conserved motifs.

Chakrabarti S(1), Bhardwaj N, Anand PA, Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences (TIFR), Bangalore 560065, India.

saikat@ncbs.res.in <saikat@ncbs.res.in>

BACKGROUND: Multiple sequence alignment algorithms are very important tools in

molecular biology today. Accurate alignment of proteins is central to several

areas such as homology modelling, docking studies, understanding evolutionary

trends and study of structure-function relationships. In recent times,

improvement of existing progressing programs and implementation of new iterative

algorithms have made a significant change in this field.

RESULTS: We report an alignment algorithm that combines progressive dynamic

algorithm, local substructure alignment and iterative refinement to achieve an

improved, user-interactive tool. Large-scale benchmarking studies show that this

FMALIGN server produces alignments that, aside from preservation of functional

and structural conservation, have accuracy comparable to other popular multiple

alignment programs.

CONCLUSIONS: The FMALIGN server allows the user to fix conserved regions in

equivalent position in the alignment thereby reducing the chance of global

misalignment to a great extent. FMALIGN is available at

http://caps.ncbs.res.in/FMALIGN/Home.html.

DOI: 10.1186/1471-2105-5-167

PMCID: PMC533867

PMID: 15509307 [Indexed for MEDLINE]

3244. BMC Bioinformatics. 2004 Oct 15;5:150.

PCOGR: phylogenetic COG ranking as an online tool to judge the specificity of

COGs with respect to freely definable groups of organisms.

Meereis F(1), Kaufmann M.

Author information:

(1)The Protein Chemistry Group, Institute of Neurobiochemistry, Witten/Herdecke

University, Stockumer Str. 10, 58448 Witten, Germany. fmeereis@uni-wh.de

<fmeereis@uni-wh.de>

BACKGROUND: The rapidly increasing number of completely sequenced genomes led to

the establishment of the COG-database which, based on sequence homologies,

assigns similar proteins from different organisms to clusters of orthologous

groups (COGs). There are several bioinformatic studies that made use of this

database to determine (hyper)thermophile-specific proteins by searching for COGs

containing (almost) exclusively proteins from (hyper)thermophilic genomes.

However, public software to perform individually definable group-specific

searches is not available.

RESULTS: The tool described here exactly fills this gap. The software is

accessible at http://www.uni-wh.de/pcogr and is linked to the COG-database. The

user can freely define two groups of organisms by selecting for each of the

(current) 66 organisms to belong either to groupA, to the reference groupB or to

be ignored by the algorithm. Then, for all COGs a specificity index is calculated

with respect to the specificity to groupA, i. e. high scoring COGs contain

proteins from the most of groupA organisms while proteins from the most organisms

assigned to groupB are absent. In addition to ranking all COGs according to the

user defined specificity criteria, a graphical visualization shows the

distribution of all COGs by displaying their abundance as a function of their

specificity indexes.

CONCLUSIONS: This software allows detecting COGs specific to a predefined group

of organisms. All COGs are ranked in the order of their specificity and a

graphical visualization allows recognizing (i) the presence and abundance of such

COGs and (ii) the phylogenetic relationship between groupA- and groupB-organisms.

The software also allows detecting putative protein-protein interactions, novel

enzymes involved in only partially known biochemical pathways, and alternate

enzymes originated by convergent evolution.

DOI: 10.1186/1471-2105-5-150

PMCID: PMC526202

PMID: 15488147 [Indexed for MEDLINE]

3245. Bioinformatics. 2004 Oct 12;20(15):2476-8. Epub 2004 Apr 15.

The SRS 3D module: integrating structures, sequences and features.

O'Donoghue SI(1), Meyer JE, Schafferhans A, Fries K.

Author information:

(1)LION bioscience AG, Waldhofer Strasse 98, 69123 Heidelberg, Germany.

contact@lionbioscience.com

In this paper we present SRS 3D, a new service that allows users to easily and

rapidly find all related structures for a given target sequence; structures can

then be viewed together with sequences, alignments and sequence features

(currently from UniProt, InterPro and PDB). Extensive user feedback confirms that

SRS 3D is intuitive and useful especially for those not expert in

structures.AVAILABILITY: An SRS 3D server is provided at http://srs3d.ebi.ac.uk/.

DOI: 10.1093/bioinformatics/bth260

PMID: 15087318 [Indexed for MEDLINE]

3246. Bioinformatics. 2004 Oct 12;20(15):2460-2. Epub 2004 Apr 8.

POLYVIEW: a flexible visualization tool for structural and functional annotations

of proteins.

Porollo AA(1), Adamczak R, Meller J.

Author information:

(1)Children's Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, OH

45229, USA.

The POLYVIEW visualization server can be used to generate protein sequence

annotations, including secondary structures, relative solvent accessibilities,

functional motifs and polymorphic sites. Two-dimensional graphical

representations in a customizable format may be generated for both known protein

structures and predictions obtained using protein structure prediction servers.

POLYVIEW may be used for automated generation of pictures with structural and

functional annotations for publications and proteomic on-line

resources.AVAILABILITY: http://polyview.cchmc.org.

DOI: 10.1093/bioinformatics/bth248

PMID: 15073023 [Indexed for MEDLINE]

3247. Bioinformatics. 2004 Oct 12;20(15):2482-4. Epub 2004 Apr 8.

The PDB-Preview database: a repository of in-silico models of 'on-hold' PDB

entries.

Fischer D(1), Paś J, Rychlewski L.

Author information:

(1)Buffalo Center of Excellence in Bioinformatics, University of Buffalo, 901

Washington Street Ste. 300, Buffalo, NY 14203, USA.

The PDB-Preview database is a dynamic web repository of in-silico predicted

three-dimensional (3D) models of experimentally determined structures that are

deposited into the PDB but are not yet publicly released, and are kept 'on-hold'.

The PDB-Preview database is automatically generated on a weekly basis by the

bioinfo.pl meta-server, which uses top-of-the-line fold-recognition methods. The

PDB-Preview provides biologists with preliminary fold assignments well before the

experimentally determined 3D structures are released.AVAILABILITY:

http://bioinfo.pl/PDB-Preview/.

DOI: 10.1093/bioinformatics/bth262

PMID: 15073011 [Indexed for MEDLINE]

3248. Bioinformatics. 2004 Oct 12;20(15):2399-400. Epub 2004 Apr 8.

PUNS: transcriptomic- and genomic-in silico PCR for enhanced primer design.

Boutros PC(1), Okey AB.

Author information:

(1)Department of Pharmacology, University of Toronto, Medical Sciences Building,

1 King's College Circle, Toronto, Ontario, M5S-1A8 Canada.

Paul.Boutros@utoronto.ca

We developed a CGI/Perl-based web server to perform in silico polymerase chain

reaction (PCR) on PCR primer sequences. The PUNS (Primer-UniGene Selectivity)

server simulates PCR reactions by running BLASTN analysis on user-entered primer

pairs against both the transcriptome and the genome to assess primer specificity.

PUNS is particularly suited for the identification of highly selective primers

for quantitative microarray validation.AVAILABILITY: Both system access and

source-code are freely available at http://okeylabimac.med.utoronto.ca/PUNS.

DOI: 10.1093/bioinformatics/bth257

PMID: 15073008 [Indexed for MEDLINE]

3249. Bioinformatics. 2004 Oct 12;20(15):2390-8. Epub 2004 Apr 8.

GEPIS--quantitative gene expression profiling in normal and cancer tissues.

Zhang Y(1), Eberhard DA, Frantz GD, Dowd P, Wu TD, Zhou Y, Watanabe C, Luoh SM,

Polakis P, Hillan KJ, Wood WI, Zhang Z.

Author information:

(1)Department of Bioinformatics, Genentech Inc., South San Francisco, CA 94080,

USA.

MOTIVATION: Expression profiling in diverse tissues is fundamental to

understanding gene function as well as therapeutic target identification. The

vast collection of expressed sequence tags (ESTs) and the associated tissue

source information provides an attractive opportunity for studying gene

expression.

RESULTS: To facilitate EST-based expression analysis, we developed GEPIS (gene

expression profiling in silico), a tool that integrates EST and tissue source

information to compute gene expression patterns in a large panel of normal and

tumor samples. We found EST-based expression patterns to be consistent with

published papers as well as our own experimental results. We also built a GEPIS

Regional Atlas that depicts expression characteristics of all genes in a selected

genomic region. This program can be adapted for large-scale screening for genes

with desirable expression patterns, as illustrated by our large-scale mining for

tissue- and tumor-specific genes.

AVAILABILITY: The email server version of the GEPIS application is freely

available at http://share.gene.com/share/gepis. An interactive version of GEPIS

will soon be freely available at

http://www.cgl.ucsf.edu/Research/genentech/gepis/. The source code, modules, data

and gene lists can be downloaded at http://share.gene.com/share/gepis.

DOI: 10.1093/bioinformatics/bth256

PMID: 15073007 [Indexed for MEDLINE]

3250. Clin Microbiol Infect. 2004 Oct;10(10):948-50.

MOP-UP: an online tool for finding strain-specific primers or motifs in DNA or

protein alignments.

Underwood AP(1), Green J.

Author information:

(1)Central Public Health Laboratory, Genomics, Proteomics and Bioinformatics

Unit, London, UK. anthony.underwood@hpa.org.uk

MOP-UP is a web-based application that enables efficient searching of nucleic

acid or amino-acid alignments for sequences or motifs that are unique to a subset

of the members represented in the alignment. This has applications in the design

of assays that aim to detect particular strains or species. Since molecular-based

characterisation of microbes is becoming increasingly important, MOP-UP can aid

microbiologists in finding the best loci on which to base such assays. The

program is accessible at:

http://www.hpa.org.uk/srmd/bioinformatics/tools/mop-ups.htm.

DOI: 10.1111/j.1469-0691.2004.00943.x

PMID: 15373897 [Indexed for MEDLINE]

3251. Comput Methods Programs Biomed. 2004 Oct;76(1):53-71.

MITIS: a WWW-based medical system for managing and processing

gynecological-obstetrical-radiological data.

Matsopoulos GK(1), Kouloulias V, Asvestas P, Mouravliansky N, Delibasis K,

Demetriades D.

Author information:

(1)Institute of Communication and Computer Systems, National Technical University

of Athens, 9, Iroon Polytechneiou Str., Zografos, Athens 157 80, Greece.

gmatso@esd.ece.ntua.gr

In this paper a World Wide Web (WWW)-based medical system, called MITIS, is

designed and developed for the management and processing of obstetrical,

gynecological and radiological medical data. The system records all the necessary

medical information in terms of patient data, examinations, and operations and

provides the user-expert with advanced image processing tools for the

manipulation, processing and storage of ultrasound and mammographic images. The

system can be installed in a hospital's Local Area Network (LAN) where it can

access picture archival and communication systems (PACS) servers (if available),

or any other server within the radiology department, for image archiving and

retrieval, based on the digital imaging and communication in medicine (DICOM) 3.0

protocol, over TCP/IP and also it is accessible to external physicians via the

hospital's Internet connection. MITIS is composed as a set of independent WWW

modules (ISAPI server extension dlls) and a Win32 application (COM+ server) for

mammography image processing and evaluation.

DOI: 10.1016/j.cmpb.2004.03.001

PMID: 15313542 [Indexed for MEDLINE]

3252. RNA. 2004 Oct;10(10):1507-17.

Fast and effective prediction of microRNA/target duplexes.

Rehmsmeier M(1), Steffen P, Hochsmann M, Giegerich R.

Author information:

(1)Universität Bielefeld, International NRW Graduate School in Bioinformatics and

Genome Research, Postfach 10 01 31, 33501 Bielefeld, Germany.

marc@techfak.uni-bielefeld.de

MicroRNAs (miRNAs) are short RNAs that post-transcriptionally regulate the

expression of target genes by binding to the target mRNAs. Although a large

number of animal miRNAs has been defined, only a few targets are known. In

contrast to plant miRNAs, which usually bind nearly perfectly to their targets,

animal miRNAs bind less tightly, with a few nucleotides being unbound, thus

producing more complex secondary structures of miRNA/target duplexes. Here, we

present a program, RNA-hybrid, that predicts multiple potential binding sites of

miRNAs in large target RNAs. In general, the program finds the energetically most

favorable hybridization sites of a small RNA in a large RNA. Intramolecular

hybridizations, that is, base pairings between target nucleotides or between

miRNA nucleotides are not allowed. For large targets, the time complexity of the

algorithm is linear in the target length, allowing many long targets to be

searched in a short time. Statistical significance of predicted targets is

assessed with an extreme value statistics of length normalized minimum free

energies, a Poisson approximation of multiple binding sites, and the calculation

of effective numbers of orthologous targets in comparative studies of multiple

organisms. We applied our method to the prediction of Drosophila miRNA targets in

3'UTRs and coding sequence. RNAhybrid, with its accompanying programs

RNAcalibrate and RNAeffective, is available for download and as a Web tool on the

Bielefeld Bioinformatics Server

(http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/).

Copyright 2004 RNA Society

DOI: 10.1261/rna.5248604

PMCID: PMC1370637

PMID: 15383676 [Indexed for MEDLINE]

3253. BMC Bioinformatics. 2004 Sep 28;5:138.

cuticleDB: a relational database of Arthropod cuticular proteins.

Magkrioti CK(1), Spyropoulos IC, Iconomidou VA, Willis JH, Hamodrakas SJ.

Author information:

(1)Department of Cell Biology and Biophysics, Faculty of Biology, University of

Athens, Athens 15701, Greece. chmagkrioti@yahoo.gr

BACKGROUND: The insect exoskeleton or cuticle is a bi-partite composite of

proteins and chitin that provides protective, skeletal and structural functions.

Little information is available about the molecular structure of this important

complex that exhibits a helicoidal architecture. Scores of sequences of cuticular

proteins have been obtained from direct protein sequencing, from cDNAs, and from

genomic analyses. Most of these cuticular protein sequences contain motifs found

only in arthropod proteins.

DESCRIPTION: cuticleDB is a relational database containing all structural

proteins of Arthropod cuticle identified to date. Many come from direct

sequencing of proteins isolated from cuticle and from sequences from cDNAs that

share common features with these authentic cuticular proteins. It also includes

proteins from the Drosophila melanogaster and the Anopheles gambiae genomes, that

have been predicted to be cuticular proteins, based on a Pfam motif (PF00379)

responsible for chitin binding in Arthropod cuticle. The total number of the

database entries is 445: 370 derive from insects, 60 from Crustacea and 15 from

Chelicerata. The database can be accessed from our web server at

http://bioinformatics.biol.uoa.gr/cuticleDB.

CONCLUSIONS: CuticleDB was primarily designed to contain correct and full

annotation of cuticular protein data. The database will be of help to future

genome annotators. Users will be able to test hypotheses for the existence of

known and also of yet unknown motifs in cuticular proteins. An analysis of motifs

may contribute to understanding how proteins contribute to the physical

properties of cuticle as well as to the precise nature of their interaction with

chitin.

DOI: 10.1186/1471-2105-5-138

PMCID: PMC522807

PMID: 15453918 [Indexed for MEDLINE]

3254. Bioinformatics. 2004 Sep 22;20(14):2222-7. Epub 2004 Apr 8.

Alignment of RNA base pairing probability matrices.

Hofacker IL(1), Bernhart SH, Stadler PF.

Author information:

(1)Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität

Wien, Währingerstrasse 17, Vienna, Austria. ivo@tbi.univie.ac.at

MOTIVATION: Many classes of functional RNA molecules are characterized by highly

conserved secondary structures but little detectable sequence similarity.

Reliable multiple alignments can therefore be constructed only when the shared

structural features are taken into account. Since multiple alignments are used as

input for many subsequent methods of data analysis, structure-based alignments

are an indispensable necessity in RNA bioinformatics.

RESULTS: We present here a method to compute pairwise and progressive multiple

alignments from the direct comparison of base pairing probability matrices.

Instead of attempting to solve the folding and the alignment problem

simultaneously as in the classical Sankoff's algorithm, we use McCaskill's

approach to compute base pairing probability matrices which effectively

incorporate the information on the energetics of each sequences. A novel,

simplified variant of Sankoff's algorithms can then be employed to extract the

maximum-weight common secondary structure and an associated alignment.

AVAILABILITY: The programs pmcomp and pmmulti described in this contribution are

implemented in Perl and can be downloaded together with the example datasets from

http://www.tbi.univie.ac.at/RNA/PMcomp/. A web server is available at

http://rna.tbi.univie.ac.at/cgi-bin/pmcgi.pl

DOI: 10.1093/bioinformatics/bth229

PMID: 15073017 [Indexed for MEDLINE]

3255. Bioinformatics. 2004 Sep 22;20(14):2317-9. Epub 2004 Apr 8.

Bellerophon: a program to detect chimeric sequences in multiple sequence

alignments.

Huber T(1), Faulkner G, Hugenholtz P.

Author information:

(1)ComBinE group, Advanced Computational Modelling Centre, The University of

Queensland, Brisbane 4072, Australia.

SUMMARY: Bellerophon is a program for detecting chimeric sequences in multiple

sequence datasets by an adaption of partial treeing analysis. Bellerophon was

specifically developed to detect 16S rRNA gene chimeras in PCR-clone libraries of

environmental samples but can be applied to other nucleotide sequence alignments.

AVAILABILITY: Bellerophon is available as an interactive web server at

http://foo.maths.uq.edu.au/~huber/bellerophon.pl

DOI: 10.1093/bioinformatics/bth226

PMID: 15073015 [Indexed for MEDLINE]

3256. Bioinformatics. 2004 Sep 22;20(14):2309-11. Epub 2004 Apr 1.

MuSiC: a tool for multiple sequence alignment with constraints.

Tsai YT(1), Huang YP, Yu CT, Lu CL.

Author information:

(1)Department of Computer Science and Information Management, Providence

University, Taiwan, ROC.

SUMMARY: MuSiC is a web server to perform the constrained alignment of a set of

sequences, such that the user-specified residues/nucleotides are aligned with

each other. The input of the MuSiC system consists of a set of protein/DNA/RNA

sequences and a set of user-specified constraints, each with a fragment of

residue/nucleotide that (approximately) appears in all input sequences. The

output of MuSiC is a constrained multiple sequence alignment in which the

fragments of the input sequences whose residues/nucleotides exhibit a given

degree of similarity to a constraint are aligned together. The current MuSiC

system is implemented in Java language and can be accessed via a simple web

interface.

AVAILABILITY: http://genome.life.nctu.edu.tw/MUSIC

DOI: 10.1093/bioinformatics/bth220

PMID: 15059840 [Indexed for MEDLINE]

3257. Bioinformatics. 2004 Sep 22;20(14):2228-35. Epub 2004 Apr 1.

Arby: automatic protein structure prediction using profile-profile alignment and

confidence measures.

von Ohsen N(1), Sommer I, Zimmer R, Lengauer T.

Author information:

(1)Institute for Algorithms and Scientific Computing (SCAI), Fraunhofer

Gesellschaft, Schloss Birlinghoven, Sankt Augustin, 53754, Germany.

Niklas.von-Oehsen@scai.fhg.de

MOTIVATION: Arby is a new server for protein structure prediction that combines

several homology-based methods for predicting the three-dimensional structure of

a protein, given its sequence. The methods used include a threading approach,

which makes use of structural information, and a profile-profile alignment

approach that incorporates secondary structure predictions. The combination of

the different methods with the help of empirically derived confidence measures

affords reliable template selection.

RESULTS: According to the recent CAFASP3 experiment, the server is one of the

most sensitive methods for predicting the structure of single domain proteins.

The quality of template selection is assessed using a fold-recognition

experiment.

AVAILABILITY: The Arby server is available through the portal of the Helmholtz

Network for Bioinformatics at http://www.hnbioinfo.de under the protein structure

category.

DOI: 10.1093/bioinformatics/bth232

PMID: 15059818 [Indexed for MEDLINE]

3258. Bioinformatics. 2004 Sep 22;20(14):2331-2. Epub 2004 Apr 1.

AliasServer: a web server to handle multiple aliases used to refer to proteins.

Iragne F(1), Barré A, Goffard N, De Daruvar A.

Author information:

(1)Centre de Bioinformatique Bordeaux, Université V. Segalen Bordeaux 2, 146 rue

Léo Saignat, 33076 Bordeaux Cedex, France.

AliasServer provides services that facilitate the assembly of data or datasets

that make use of different identifiers for refering to the same protein. This

resource relies on a database which contains, for a given organism, a

non-redundant list of protein sequences associated with a set of

aliases.AVAILABILITY: AliasServer is available as an interactive Web server at

http://cbi.labri.fr/outils/alias/ and as a web service using a SOAP interface.

The complete tool, including sources and data, is available for local

installations upon request.

SUPPLEMENTARY INFORMATION: Technical documentation is available at

http://cbi.labri.fr/outils/alias/asdoc.pdf

DOI: 10.1093/bioinformatics/bth241

PMID: 15059813 [Indexed for MEDLINE]

3259. BMC Bioinformatics. 2004 Sep 22;5:135.

Profiled support vector machines for antisense oligonucleotide efficacy

prediction.

Camps-Valls G(1), Chalk AM, Serrano-López AJ, Martín-Guerrero JD, Sonnhammer EL.

Author information:

(1)Grup de Processament Digital de Senyals, Universitat de València, Spain, C/

Dr, Moliner, 50, 46100 Burjassot, València, Spain. gustavo.camps@uv.es

BACKGROUND: This paper presents the use of Support Vector Machines (SVMs) for

prediction and analysis of antisense oligonucleotide (AO) efficacy. The collected

database comprises 315 AO molecules including 68 features each, inducing a

problem well-suited to SVMs. The task of feature selection is crucial given the

presence of noisy or redundant features, and the well-known problem of the curse

of dimensionality. We propose a two-stage strategy to develop an optimal model:

(1) feature selection using correlation analysis, mutual information, and

SVM-based recursive feature elimination (SVM-RFE), and (2) AO prediction using

standard and profiled SVM formulations. A profiled SVM gives different weights to

different parts of the training data to focus the training on the most important

regions.

RESULTS: In the first stage, the SVM-RFE technique was most efficient and robust

in the presence of low number of samples and high input space dimension. This

method yielded an optimal subset of 14 representative features, which were all

related to energy and sequence motifs. The second stage evaluated the performance

of the predictors (overall correlation coefficient between observed and predicted

efficacy, r; mean error, ME; and root-mean-square-error, RMSE) using 8-fold and

minus-one-RNA cross-validation methods. The profiled SVM produced the best

results (r = 0.44, ME = 0.022, and RMSE= 0.278) and predicted high (>75%

inhibition of gene expression) and low efficacy (<25%) AOs with a success rate of

83.3% and 82.9%, respectively, which is better than by previous approaches. A web

server for AO prediction is available online at http://aosvm.cgb.ki.se/.

CONCLUSIONS: The SVM approach is well suited to the AO prediction problem, and

yields a prediction accuracy superior to previous methods. The profiled SVM was

found to perform better than the standard SVM, suggesting that it could lead to

improvements in other prediction problems as well.

DOI: 10.1186/1471-2105-5-135

PMCID: PMC526382

PMID: 15383156 [Indexed for MEDLINE]

3260. BMC Bioinformatics. 2004 Sep 21;5:134.

PATTERNFINDER: combined analysis of DNA regulatory sequences and double-helix

stability.

Huang Y(1), Kowalski D.

Author information:

(1)Cancer Genetics Department, Roswell Park Cancer Institute, Buffalo, NY 14263,

USA. yanlinh@microsoft.com

BACKGROUND: Regulatory regions that function in DNA replication and gene

transcription contain specific sequences that bind proteins as well as

less-specific sequences in which the double helix is often easy to unwind.

Progress towards predicting and characterizing regulatory regions could be

accelerated by computer programs that perform a combined analysis of specific

sequences and DNA unwinding properties.

RESULTS: Here we present PATTERNFINDER, a web server that searches DNA sequences

for matches to specific or flexible patterns, and analyzes DNA helical stability.

A batch mode of the program generates a tabular map of matches to multiple,

different patterns. Regions flanking pattern matches can be targeted for helical

stability analysis to identify sequences with a minimum free energy for DNA

unwinding. As an example application, we analyzed a regulatory region of the

human c-myc proto-oncogene consisting of a single-strand-specific protein binding

site within a DNA region that unwindsin vivo. The predicted region of minimal

helical stability overlapped both the protein binding site and the unwound DNA

region identified experimentally.

CONCLUSIONS: The PATTERNFINDER web server permits localization of known

functional elements or landmarks in DNA sequences as well as prediction of

potential new elements. Batch analysis of multiple patterns facilitates the

annotation of DNA regulatory regions. Identifying specific pattern matches linked

to DNA with low helical stability is useful in characterizing regulatory regions

for transcription, replication and other processes and may predict functional DNA

unwinding elements.PATTERNFINDER can be accessed freely at:

http://wings.buffalo.edu/gsa/dna/dk/PFP/

DOI: 10.1186/1471-2105-5-134

PMCID: PMC520813

PMID: 15383143 [Indexed for MEDLINE]

3261. BMC Bioinformatics. 2004 Sep 9;5:130.

GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae

genomics and genetics research.

Jung S(1), Jesudurai C, Staton M, Du Z, Ficklin S, Cho I, Abbott A, Tomkins J,

Main D.

Author information:

(1)Department of Genetics and Biochemistry, Clemson University, Clemson, SC,

29634, USA. sook@genome.clemson.edu

BACKGROUND: Peach is being developed as a model organism for Rosaceae, an

economically important family that includes fruits and ornamental plants such as

apple, pear, strawberry, cherry, almond and rose. The genomics and genetics data

of peach can play a significant role in the gene discovery and the genetic

understanding of related species. The effective utilization of these peach

resources, however, requires the development of an integrated and centralized

database with associated analysis tools.

DESCRIPTION: The Genome Database for Rosaceae (GDR) is a curated and integrated

web-based relational database. GDR contains comprehensive data of the genetically

anchored peach physical map, an annotated peach EST database, Rosaceae maps and

markers and all publicly available Rosaceae sequences. Annotations of ESTs

include contig assembly, putative function, simple sequence repeats, and anchored

position to the peach physical map where applicable. Our integrated map viewer

provides graphical interface to the genetic, transcriptome and physical mapping

information. ESTs, BACs and markers can be queried by various categories and the

search result sites are linked to the integrated map viewer or to the WebFPC

physical map sites. In addition to browsing and querying the database, users can

compare their sequences with the annotated GDR sequences via a dedicated sequence

similarity server running either the BLAST or FASTA algorithm. To demonstrate the

utility of the integrated and fully annotated database and analysis tools, we

describe a case study where we anchored Rosaceae sequences to the peach physical

and genetic map by sequence similarity.

CONCLUSIONS: The GDR has been initiated to meet the major deficiency in Rosaceae

genomics and genetics research, namely a centralized web database and

bioinformatics tools for data storage, analysis and exchange. GDR can be accessed

at http://www.genome.clemson.edu/gdr/.

DOI: 10.1186/1471-2105-5-130

PMCID: PMC517928

PMID: 15357877 [Indexed for MEDLINE]

3262. Bioinformatics. 2004 Sep 1;20(13):2143-4. Epub 2004 Apr 8.

Web-based kinetic modelling using JWS Online.

Olivier BG(1), Snoep JL.

Author information:

(1)Department of Biochemistry, University of Stellenbosch, Private Bag X1,

Matieland 7602, South Africa.

JWS Online is a repository of kinetic models, describing biological systems,

which can be interactively run and interrogated over the Internet. It is

implemented using a client-server strategy where the clients, in the form of web

browser based Java applets, act as a graphical interface to the model servers,

which perform the required numerical computations.AVAILABILITY: The JWS Online

website is publicly accessible at http://jjj.biochem.sun.ac.za/ with mirrors at

http://www.jjj.bio.vu.nl/ and http://jjj.vbi.vt.edu/

DOI: 10.1093/bioinformatics/bth200

PMID: 15072998 [Indexed for MEDLINE]

3263. Bioinformatics. 2004 Sep 1;20(13):2150-2. Epub 2004 Apr 1.

F2CS: FSSP to CATH and SCOP prediction server.

Getz G(1), Starovolsky A, Domany E.

Author information:

(1)Department of Physics of Complex Systems, Weizmann Institute of Science,

Rehovot 76100, Israel.

The F2CS server provides access to the software, F2CS2.00, which implements an

automated prediction method of SCOP and CATH classifications of proteins, based

on their FSSP Z-scores.AVAILABILITY: Free at

http://www.weizmann.ac.il/physics/complex/compphys/f2cs/

SUPPLEMENTARY INFORMATION: The site contains links to additional figures and

tables.

DOI: 10.1093/bioinformatics/bth208

PMID: 15059833 [Indexed for MEDLINE]

3264. Bioinformatics. 2004 Sep 1;20(13):2084-91. Epub 2004 Apr 1.

Gene annotation from scientific literature using mappings between keyword

systems.

Pérez AJ(1), Perez-Iratxeta C, Bork P, Thode G, Andrade MA.

Author information:

(1)University of Málaga, Facultad de Ciencias, Departmento de Genetica, Group of

Bioinformatics, Campus Universitario de Teatinos, 29071 Málaga, Spain.

MOTIVATION: The description of genes in databases by keywords helps the

non-specialist to quickly grasp the properties of a gene and increases the

efficiency of computational tools that are applied to gene data (e.g. searching a

gene database for sequences related to a particular biological process). However,

the association of keywords to genes or protein sequences is a difficult process

that ultimately implies examination of the literature related to a gene.

RESULTS: To support this task, we present a procedure to derive keywords from the

set of scientific abstracts related to a gene. Our system is based on the

automated extraction of mappings between related terms from different databases

using a model of fuzzy associations that can be applied with all generality to

any pair of linked databases. We tested the system by annotating genes of the

SWISS-PROT database with keywords derived from the abstracts linked to their

entries (stored in the MEDLINE database of scientific references). The

performance of the annotation procedure was much better for SWISS-PROT keywords

(recall of 47%, precision of 68%) than for Gene Ontology terms (recall of 8%,

precision of 67%).

AVAILABILITY: The algorithm can be publicly accessed and used for the annotation

of sequences through a web server at http://www.bork.embl.de/kat

DOI: 10.1093/bioinformatics/bth207

PMID: 15059832 [Indexed for MEDLINE]

3265. Bioinformatics. 2004 Sep 1;20(13):2138-9. Epub 2004 Mar 25.

The DISOPRED server for the prediction of protein disorder.

Ward JJ(1), McGuffin LJ, Bryson K, Buxton BF, Jones DT.

Author information:

(1)Bioinformatics Group, Department of Computer Science, University College

London, Gower Street, London, WC1E 6BT, UK.

Dynamically disordered regions appear to be relatively abundant in eukaryotic

proteomes. The DISOPRED server allows users to submit a protein sequence, and

returns a probability estimate of each residue in the sequence being disordered.

The results are sent in both plain text and graphical formats, and the server can

also supply predictions of secondary structure to provide further structural

information.AVAILABILITY: The server can be accessed by non-commercial users at

http://bioinf.cs.ucl.ac.uk/disopred/

DOI: 10.1093/bioinformatics/bth195

PMID: 15044227 [Indexed for MEDLINE]

3266. Genome Res. 2004 Sep;14(9):1756-66.

EGPred: prediction of eukaryotic genes using ab initio methods after combining

with sequence similarity approaches.

Issac B(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector 39A, Chandigarh-160036. India.

EGPred is a Web-based server that combines ab initio methods and similarity

searches to predict genes, particularly exon regions, with high accuracy. The

EGPred program proceeds in the following steps: (1) an initial BLASTX search of

genomic sequence against the RefSeq database is used to identify protein hits

with an E-value <1; (2) a second BLASTX search of genomic sequence against the

hits from the previous run with relaxed parameters (E-values <10) helps to

retrieve all probable coding exon regions; (3) a BLASTN search of genomic

sequence against the intron database is then used to detect probable intron

regions; (4) the probable intron and exon regions are compared to filter/remove

wrong exons; (5) the NNSPLICE program is then used to reassign splicing signal

site positions in the remaining probable coding exons; and (6) finally ab initio

predictions are combined with exons derived from the fifth step based on the

relative strength of start/stop and splice signal sites as obtained from ab

initio and similarity search. The combination method increases the exon level

performance of five different ab initio programs by 4%-10% when evaluated on the

HMR195 data set. Similar improvement is observed when ab initio programs are

evaluated on the Burset/Guigo data set. Finally, EGPred is demonstrated on an

approximately 95-Mbp fragment of human chromosome 13. The list of predicted genes

from this analysis are available in the supplementary material. The EGPred

program is computationally intensive due to multiple BLAST runs during each

analysis. The EGPred server is available at

http://www.imtech.res.in/raghava/egpred/.

DOI: 10.1101/gr.2524704

PMCID: PMC515322

PMID: 15342559 [Indexed for MEDLINE]

3267. Proteins. 2004 Sep 1;56(4):753-67.

Accurate prediction of solvent accessibility using neural networks-based

regression.

Adamczak R(1), Porollo A, Meller J.

Author information:

(1)Children's Hospital Research Foundation, Cincinnati, Ohio, USA.

Accurate prediction of relative solvent accessibilities (RSAs) of amino acid

residues in proteins may be used to facilitate protein structure prediction and

functional annotation. Toward that goal we developed a novel method for improved

prediction of RSAs. Contrary to other machine learning-based methods from the

literature, we do not impose a classification problem with arbitrary boundaries

between the classes. Instead, we seek a continuous approximation of the

real-value RSA using nonlinear regression, with several feed forward and

recurrent neural networks, which are then combined into a consensus predictor. A

set of 860 protein structures derived from the PFAM database was used for

training, whereas validation of the results was carefully performed on several

nonredundant control sets comprising a total of 603 structures derived from new

Protein Data Bank structures and had no homology to proteins included in the

training. Two classes of alternative predictors were developed for comparison

with the regression-based approach: one based on the standard classification

approach and the other based on a semicontinuous approximation with the so-called

thermometer encoding. Furthermore, a weighted approximation, with errors being

scaled by the observed levels of variability in RSA for equivalent residues in

families of homologous structures, was applied in order to improve the results.

The effects of including evolutionary profiles and the growth of sequence

databases were assessed. In accord with the observed levels of variability in RSA

for different ranges of RSA values, the regression accuracy is higher for buried

than for exposed residues, with overall 15.3-15.8% mean absolute errors and

correlation coefficients between the predicted and experimental values of

0.64-0.67 on different control sets. The new method outperforms

classification-based algorithms when the real value predictions are projected

onto two-class classification problems with several commonly used thresholds to

separate exposed and buried residues. For example, classification accuracy of

about 77% is consistently achieved on all control sets with a threshold of 25%

RSA. A web server that enables RSA prediction using the new method and provides

customizable graphical representation of the results is available at

http://sable.cchmc.org.

Copyright 2004 Wiley-Liss, Inc.

DOI: 10.1002/prot.20176

PMID: 15281128 [Indexed for MEDLINE]

3268. Bioinformatics. 2004 Aug 12;20(12):1968-70. Epub 2004 Mar 22.

BLAST2GENE: a comprehensive conversion of BLAST output into independent genes and

gene fragments.

Suyama M(1), Torrents D, Bork P.

Author information:

(1)Biocomputing, European Molecular Biology Laboratory, Meyerhofstrasse 1,

D-69012 Heidelberg, Germany.

SUMMARY: BLAST2GENE is a program that allows a detailed analysis of genomic

regions containing completely or partially duplicated genes. From a BLAST (or

BL2SEQ) comparison of a protein or nucleotide query sequence with any genomic

region of interest, BLAST2GENE processes all high scoring pairwise alignments

(HSPs) and provides the disposition of all independent copies along the genomic

fragment. The results are provided in text and PostScript formats to allow an

automatic and visual evaluation of the respective region.

AVAILABILITY: The program is available upon request from the authors. A web

server of BLAST2GENE is maintained at http://www.bork.embl.de/blast2gene

DOI: 10.1093/bioinformatics/bth225

PMID: 15037510 [Indexed for MEDLINE]

3269. Bioinformatics. 2004 Aug 12;20(12):1822-35. Epub 2004 Feb 26.

A knowledge-based scale for the analysis and prediction of buried and exposed

faces of transmembrane domain proteins.

Beuming T(1), Weinstein H.

Author information:

(1)Department of Physiology and Biophysics, Mount Sinai School of Medicine, New

York, NY 10029, USA.

MOTIVATION: The dearth of structural data on alpha-helical membrane proteins

(MPs) has hampered thus far the development of reliable knowledge-based

potentials that can be used for automatic prediction of transmembrane (TM)

protein structure. While algorithms for identifying TM segments are available,

modeling of the TM domains of alpha-helical MPs involves assembling the segments

into a bundle. This requires the correct assignment of the buried and

lipid-exposed faces of the TM domains.

RESULTS: A recent increase in the number of crystal structures of alpha-helical

MPs has enabled an analysis of the lipid-exposed surfaces and the interiors of

such molecules on the basis of structure, rather than sequence alone. Together

with a conservation criterion that is based on previous observations that

conserved residues are mostly found in the interior of MPs, the bias of certain

residue types to be preferably buried or exposed is proposed as a criterion for

predicting the lipid-exposed and interior faces of TMs. Applications to known

structures demonstrates 80% accuracy of this prediction algorithm.

AVAILABILITY: The algorithm used for the predictions is implemented in the

ProperTM Web server (http://icb.med.cornell.edu/services/propertm/start).

DOI: 10.1093/bioinformatics/bth143

PMID: 14988128 [Indexed for MEDLINE]

3270. Bioinformatics. 2004 Aug 12;20(12):1850-60. Epub 2004 Feb 26.

Multiple-sequence functional annotation and the generalized hidden Markov

phylogeny.

McAuliffe JD(1), Pachter L, Jordan MI.

Author information:

(1)Department of Statistics, University of California, 367 Evans Hall, Berkeley,

CA 94720, USA.

MOTIVATION: Phylogenetic shadowing is a comparative genomics principle that

allows for the discovery of conserved regions in sequences from multiple closely

related organisms. We develop a formal probabilistic framework for combining

phylogenetic shadowing with feature-based functional annotation methods. The

resulting model, a generalized hidden Markov phylogeny (GHMP), applies to a

variety of situations where functional regions are to be inferred from

evolutionary constraints.

RESULTS: We show how GHMPs can be used to predict complete shared gene structures

in multiple primate sequences. We also describe shadower, our implementation of

such a prediction system. We find that shadower outperforms previously reported

ab initio gene finders, including comparative human-mouse approaches, on a small

sample of diverse exonic regions. Finally, we report on an empirical analysis of

shadower's performance which reveals that as few as five well-chosen species may

suffice to attain maximal sensitivity and specificity in exon demarcation.

AVAILABILITY: A Web server is available at http://bonaire.lbl.gov/shadower

DOI: 10.1093/bioinformatics/bth153

PMID: 14988105 [Indexed for MEDLINE]

3271. Bioinformatics. 2004 Aug 12;20(12):1842-9. Epub 2004 Feb 26.

Haplotype reconstruction from genotype data using Imperfect Phylogeny.

Halperin E(1), Eskin E.

Author information:

(1)CS Division, University of California Berkeley, Berkeley, CA 92093-0114, USA.

eran@eecs.berkeley.edu

Critical to the understanding of the genetic basis for complex diseases is the

modeling of human variation. Most of this variation can be characterized by

single nucleotide polymorphisms (SNPs) which are mutations at a single nucleotide

position. To characterize the genetic variation between different people, we must

determine an individual's haplotype or which nucleotide base occurs at each

position of these common SNPs for each chromosome. In this paper, we present

results for a highly accurate method for haplotype resolution from genotype data.

Our method leverages a new insight into the underlying structure of haplotypes

that shows that SNPs are organized in highly correlated 'blocks'. In a few recent

studies, considerable parts of the human genome were partitioned into blocks,

such that the majority of the sequenced genotypes have one of about four common

haplotypes in each block. Our method partitions the SNPs into blocks, and for

each block, we predict the common haplotypes and each individual's haplotype. We

evaluate our method over biological data. Our method predicts the common

haplotypes perfectly and has a very low error rate (<2% over the data) when

taking into account the predictions for the uncommon haplotypes. Our method is

extremely efficient compared with previous methods such as PHASE and HAPLOTYPER.

Its efficiency allows us to find the block partition of the haplotypes, to cope

with missing data and to work with large datasets.AVAILABILITY: The algorithm is

available via a Web server at http://www.calit2.net/compbio/hap/

DOI: 10.1093/bioinformatics/bth149

PMID: 14988101 [Indexed for MEDLINE]

3272. Bioinformatics. 2004 Aug 4;20 Suppl 1:i311-7.

Tracking repeats using significance and transitivity.

Szklarczyk R(1), Heringa J.

Author information:

(1)Centre for Integrative Bioinformatics (IBIVU), Faculty of Sciences and Faculty

of Earth and Life Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1081A,

Amsterdam, The Netherlands. radek@cs.vu.nl

MOTIVATION: Internal repeats in coding sequences correspond to structural and

functional units of proteins. Moreover, duplication of fragments of coding

sequences is known to be a mechanism to facilitate evolution. Identification of

repeats is crucial to shed light on the function and structure of proteins, and

explain their evolutionary past. The task is difficult because during the course

of evolution many repeats diverged beyond recognition.

RESULTS: We introduce a new method TRUST, for ab initio determination of internal

repeats in proteins. It provides an improvement in prediction quality as compared

to alternative state-of-the-art methods. The increased sensitivity and accuracy

of the method is achieved by exploiting the concept of transitivity of

alignments. Starting from significant local suboptimal alignments, the

application of transitivity allows us to (1) identify distant repeat homologues

for which no alignments were found; (2) gain confidence about consistently

well-aligned regions; and (3) recognize and reduce the contribution of

non-homologous repeats. This re-assessment step enables us to derive a virtually

noise-free profile representing a generalized repeat with high fidelity. We also

obtained superior specificity by employing rigid statistical testing for

self-sequence and profile-sequence alignments. Assessment was done using a

database of repeat annotations based on structural superpositioning. The results

show that TRUST is a useful and reliable tool for mining tandem and non-tandem

repeats in protein sequence databases, capable of predicting multiple repeat

types with varying intervening segments within a single sequence.

AVAILABILITY: The TRUST server (together with the source code) is available at

http://ibivu.cs.vu.nl/programs/trustwww

DOI: 10.1093/bioinformatics/bth911

PMID: 15262814 [Indexed for MEDLINE]

3273. Bioinformatics. 2004 Aug 4;20 Suppl 1:i63-8.

A neural-network-based method for predicting protein stability changes upon

single point mutations.

Capriotti E(1), Fariselli P, Casadio R.

Author information:

(1)Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna,

Bologna, Italy.

MOTIVATION: One important requirement for protein design is to be able to predict

changes of protein stability upon mutation. Different methods addressing this

task have been described and their performance tested considering global linear

correlation between predicted and experimental data. Neither is direct

statistical evaluation of their prediction performance available, nor is a direct

comparison among different approaches possible. Recently, a significant database

of thermodynamic data on protein stability changes upon single point mutation has

been generated (ProTherm). This allows the application of machine learning

techniques to predicting free energy stability changes upon mutation starting

from the protein sequence.

RESULTS: In this paper, we present a neural-network-based method to predict if a

given mutation increases or decreases the protein thermodynamic stability with

respect to the native structure. Using a dataset consisting of 1615 mutations,

our predictor correctly classifies >80% of the mutations in the database. On the

same task and using the same data, our predictor performs better than other

methods available on the Web. Moreover, when our system is coupled with

energy-based methods, the joint prediction accuracy increases up to 90%,

suggesting that it can be used to increase also the performance of pre-existing

methods, and generally to improve protein design strategies.

AVAILABILITY: The server is under construction and will be available at

http://www.biocomp.unibo.it

DOI: 10.1093/bioinformatics/bth928

PMID: 15262782 [Indexed for MEDLINE]

3274. AJR Am J Roentgenol. 2004 Aug;183(2):535-7.

MyFreePACS: a free web-based radiology image storage and viewing tool.

de Regt D(1), Weinberger E.

Author information:

(1)Reed College, 3203 SE Woodstock Blvd., Portland, OR 97202-8199, USA.

OBJECTIVE: We developed an easy-to-use method for central storage and subsequent

viewing of radiology images for use on any PC equipped with Internet Explorer.

CONCLUSION: We developed MyFreePACS, a program that uses a DICOM server to

receive and store images and transmit them over the Web to the MyFreePACS Web

client. The MyFreePACS Web client is a Web page that uses an ActiveX control for

viewing and manipulating images. The client contains many of the tools found in

modern image viewing stations including 3D localization and multiplanar

reformation. The system is built entirely with free components and is freely

available for download and installation from the Web at www.myfreepacs.com.

DOI: 10.2214/ajr.183.2.1830535

PMID: 15269053 [Indexed for MEDLINE]

3275. BMC Genomics. 2004 Aug 1;5(1):51.

C-type lectin-like domains in Fugu rubripes.

Zelensky AN(1), Gready JE.

Author information:

(1)Computational Proteomics and Therapy Design Group, John Curtin School of

Medical Research, Australian National University, PO Box 334, Canberra, ACT 2601,

Australia. Alex.Zelensky@anu.edu.au

BACKGROUND: Members of the C-type lectin domain (CTLD) superfamily are metazoan

proteins functionally important in glycoprotein metabolism, mechanisms of

multicellular integration and immunity. Three genome-level studies on human, C.

elegans and D. melanogaster reported previously demonstrated almost complete

divergence among invertebrate and mammalian families of CTLD-containing proteins

(CTLDcps).

RESULTS: We have performed an analysis of CTLD family composition in Fugu

rubripes using the draft genome sequence. The results show that all but two

groups of CTLDcps identified in mammals are also found in fish, and that most of

the groups have the same members as in mammals. We failed to detect

representatives for CTLD groups V (NK cell receptors) and VII (lithostathine),

while the DC-SIGN subgroup of group II is overrepresented in Fugu. Several new

CTLD-containing genes, highly conserved between Fugu and human, were discovered

using the Fugu genome sequence as a reference, including a CSPG family member and

an SCP-domain-containing soluble protein. A distinct group of soluble dual-CTLD

proteins has been identified, which may be the first reported CTLDcp group shared

by invertebrates and vertebrates. We show that CTLDcp-encoding genes are

selectively duplicated in Fugu, in a manner that suggests an ancient large-scale

duplication event. We have verified 32 gene structures and predicted 63 new ones,

and make our annotations available through a distributed annotation system (DAS)

server http://anz.anu.edu.au:8080/Fugu\_rubripes/ and their sequences as

additional files with this paper.

CONCLUSIONS: The vertebrate CTLDcp family was essentially formed early in

vertebrate evolution and is completely different from the invertebrate families.

Comparison of fish and mammalian genomes revealed three groups of CTLDcps and

several new members of the known groups, which are highly conserved between fish

and mammals, but were not identified in the study using only mammalian genomes.

Despite limitations of the draft sequence, the Fugu rubripes genome is a powerful

instrument for gene discovery and vertebrate evolutionary analysis. The

composition of the CTLDcp superfamily in fish and mammals suggests that

large-scale duplication events played an important role in the evolution of

vertebrates.

DOI: 10.1186/1471-2164-5-51

PMCID: PMC514892

PMID: 15285787 [Indexed for MEDLINE]

3276. Mech Ageing Dev. 2004 Aug;125(8):547-52.

Developments of geriatric autopsy database and Internet-based database of

Japanese single nucleotide polymorphisms for geriatric research (JG-SNP).

Sawabe M(1), Arai T, Kasahara I, Esaki Y, Nakahara K, Hosoi T, Orimo H, Takubo K,

Murayama S, Tanaka N; Tokyo Metropolitan Geriatric Medical Center; Japan Science

and Technology Agency.

Author information:

(1)Department of Pathology, Tokyo Metropolitan Geriatric Medical Center, 35-2

Sakae-cho, Itabashi, Tokyo 173-0015, Japan. sawabe@tmig.or.jp

To facilitate geriatric research on the roles of genetic polymorphisms of

candidate genes, two databases were developed based on data obtained from autopsy

examinations of elderly subjects: the geriatric autopsy database (GEAD) and the

Japanese single nucleotide polymorphisms (SNP) database for geriatric research

(JG-SNP) which is accessible on the Internet

(http://www.tmgh.metro.tokyo.jp/jg-snp/english/E\_top.html). The data for the GEAD

were derived from 1074 consecutive autopsy cases (565 male and 509 female cases)

with an average age of 80 years. The GEAD was installed on a stand-alone Windows

2000 server using Oracle 8i as the database application. The GEAD contains

clinical diagnoses of 26 geriatric diseases, histories of smoking and alcohol

consumption, pathological findings (720 items), severity of atherosclerosis,

genetic polymorphism data, etc. On the JG-SNP website, case distribution

corresponding to a specified SNP or disease can be searched or downloaded.

Although there are several Internet-based SNP databases such as dbSNP, no

databases are available at present on the web that contain both SNP data and

phenotypic data. As autopsy studies can provide large amounts of accurate medical

information, including the presence of undiagnosed diseases such as latent

cancers, the GEAD is a unique and excellent database for research on genetic

polymorphisms.

DOI: 10.1016/j.mad.2004.06.005

PMID: 15336912 [Indexed for MEDLINE]

3277. BMC Bioinformatics. 2004 Jul 22;5:98.

Integrated web service for improving alignment quality based on segments

comparison.

Plewczynski D(1), Rychlewski L, Ye Y, Jaroszewski L, Godzik A.

Author information:

(1)Bioinformatics Laboratory, BioInfoBank Institute, Poznan, Poland.

darman@bioinfo.pl

BACKGROUND: Defining blocks forming the global protein structure on the basis of

local structural regularity is a very fruitful idea, extensively used in

description, and prediction of structure from only sequence information. Over

many years the secondary structure elements were used as available building

blocks with great success. Specially prepared sets of possible structural motifs

can be used to describe similarity between very distant, non-homologous proteins.

The reason for utilizing the structural information in the description of

proteins is straightforward. Structural comparison is able to detect

approximately twice as many distant relationships as sequence comparison at the

same error rate.

RESULTS: Here we provide a new fragment library for Local Structure Segment (LSS)

prediction called FRAGlib which is integrated with a previously described segment

alignment algorithm SEA. A joined FRAGlib/SEA server provides easy access to both

algorithms, allowing a one stop alignment service using a novel approach to

protein sequence alignment based on a network matching approach. The FRAGlib used

as secondary structure prediction achieves only 73% accuracy in Q3 measure, but

when combined with the SEA alignment, it achieves a significant improvement in

pairwise sequence alignment quality, as compared to previous SEA implementation

and other public alignment algorithms. The FRAGlib algorithm takes approximately

2 min. to search over FRAGlib database for a typical query protein with 500

residues. The SEA service align two typical proteins within circa approximately 5

min. All supplementary materials (detailed results of all the benchmarks, the

list of test proteins and the whole fragments library) are available for download

on-line at http://ffas.ljcrf.edu/darman/results/.

CONCLUSIONS: The joined FRAGlib/SEA server will be a valuable tool both for

molecular biologists working on protein sequence analysis and for

bioinformaticians developing computational methods of structure prediction and

alignment of proteins.

DOI: 10.1186/1471-2105-5-98

PMCID: PMC497040

PMID: 15271224 [Indexed for MEDLINE]

3278. Bioinformatics. 2004 Jul 10;20(10):1641-3. Epub 2004 Feb 12.

VARAN: a web server for variability analysis of DNA microarray experiments.

Golfier G(1), Dang MT, Dauphinot L, Graison E, Rossier J, Potier MC.

Author information:

(1)Neurobiologie et Diversité Cellulaire, CNRS UMR7637, Ecole Supérieure de

Physique et de Chimie Industrielles, 10 rue Vauquelin, 75005 Paris, France.

geoffroy.golfier@espci.fr

Here, we describe a tool for VARiability Analysis of DNA microarrays experiments

(VARAN), a freely available Web server that performs a signal intensity based

analysis of the log2 expression ratio variability deduced from DNA microarray

data (one or two channels). Two modules are proposed: VARAN generator to compute

a sliding windows analysis of the experimental variability (mean and SD) and

VARAN analyzer to compare experimental data with an asymptotic variability model

previously built with the generator module from control experiments. Both modules

provide normalized intensity signals with five possible methods, log ratio values

and a list of genes showing significant variations between

conditions.AVAILABILITY: http://www.bionet.espci.fr/varan/

SUPPLEMENTARY INFORMATION: http://www.bionet.espci.fr/varan/help.html

DOI: 10.1093/bioinformatics/bth117

PMID: 14962915 [Indexed for MEDLINE]

3279. Appl Environ Microbiol. 2004 Jul;70(7):4390-2.

Online tool for analysis of denaturing gradient gel electrophoresis profiles.

Huber F(1), Peduzzi P.

Author information:

(1)Institute of Ecology and Conservation Biology, University of Vienna, Vienna,

Austria.

We present an online tool (EquiBands,

http://www.univie.ac.at/IECB/limno/equibands/EquiBands.html) that quantifies the

matching of two bands considered to be the same in different samples, even when

samples are applied to different denaturing gradient gel electrophoresis gels.

With an environmental example we demonstrate the procedure for the classification

of two bands of different samples with the help of EquiBands.

DOI: 10.1128/AEM.70.7.4390-4392.2004

PMCID: PMC444777

PMID: 15240327 [Indexed for MEDLINE]

3280. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W679-85.

MoViES: molecular vibrations evaluation server for analysis of fluctuational

dynamics of proteins and nucleic acids.

Cao ZW(1), Xue Y, Han LY, Xie B, Zhou H, Zheng CJ, Lin HH, Chen YZ.

Author information:

(1)Department of Computational Science, National University of Singapore, Blk

SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore.

Analysis of vibrational motions and thermal fluctuational dynamics is a widely

used approach for studying structural, dynamic and functional properties of

proteins and nucleic acids. Development of a freely accessible web server for

computation of vibrational and thermal fluctuational dynamics of biomolecules is

thus useful for facilitating the relevant studies. We have developed a computer

program for computing vibrational normal modes and thermal fluctuational

properties of proteins and nucleic acids and applied it in several studies. In

our program, vibrational normal modes are computed by using modified AMBER

molecular mechanics force fields, and thermal fluctuational properties are

computed by means of a self-consistent harmonic approximation method. A web

version of our program, MoViES (Molecular Vibrations Evaluation Server), was set

up to facilitate the use of our program to study vibrational dynamics of proteins

and nucleic acids. This software was tested on selected proteins, which show that

the computed normal modes and thermal fluctuational bond disruption probabilities

are consistent with experimental findings and other normal mode computations.

MoViES can be accessed at http://ang.cz3.nus.edu.sg/cgi-bin/prog/norm.pl.

DOI: 10.1093/nar/gkh384

PMCID: PMC441522

PMID: 15215475 [Indexed for MEDLINE]

3281. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W674-8.

Isotopica: a tool for the calculation and viewing of complex isotopic envelopes.

Fernandez-de-Cossio J(1), Gonzalez LJ, Satomi Y, Betancourt L, Ramos Y, Huerta V,

Amaro A, Besada V, Padron G, Minamino N, Takao T.

Author information:

(1)Center for Genetic Engineering and Biotechnology, P.O. Box 6162, Havana, Cuba.

jorge.cossio@cigb.edu.cu

Erratum in

Nucleic Acids Res. 2004 Jul 15;32(12):3779.

The web application Isotopica has been developed as an aid to the interpretation

of ions that contain naturally occurring isotopes in a mass spectrum. It allows

the calculation of mass values and isotopic distributions based on molecular

formulas, peptides/proteins, DNA/RNA, carbohydrate sequences or combinations

thereof. In addition, Isotopica takes modifications of the input molecule into

consideration using a simple and flexible language as a straightforward extension

of the molecular formula syntax. This function is especially useful for

biomolecules, which are often subjected to additional modifications other than

normal constituents, such as the frequently occurring post-translational

modification in proteins. The isotopic distribution of any molecule thus defined

can be calculated by considering full widths at half maximum or mass resolution.

The combined envelope of several overlapping isotopic distributions of a mixture

of molecules can be determined after specifying each molecule's relative

abundance. The results can be displayed graphically on a local PC using the

Isotopica viewer, a standalone application that is downloadable from the sites

below, as a complement to the client browser. The m/z and intensity values can

also be obtained in the form of a plain ASCII text file. The software has proved

to be useful for peptide mass fingerprinting and validating an observed isotopic

ion distribution with reference to the theoretical one, even from a

multi-component sample. The web server can be accessed at

http://bioinformatica.cigb.edu.cu/isotopica and

http://coco.protein.osaka-u.ac.jp/isotopica [correction].

DOI: 10.1093/nar/gkh423

PMCID: PMC441561

PMID: 15215474 [Indexed for MEDLINE]

3282. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W668-73.

DICHROWEB, an online server for protein secondary structure analyses from

circular dichroism spectroscopic data.

Whitmore L(1), Wallace BA.

Author information:

(1)Department of Crystallography, Birkbeck College, University of London, London

WC1E 7HX, UK.

The DICHROWEB web server enables on-line analyses of circular dichroism (CD)

spectroscopic data, providing calculated secondary structure content and

graphical analyses comparing calculated structures and experimental data. The

server is located at http://www.cryst.bbk.ac.uk/cdweb and may be accessed via a

password-limited user ID, available upon completion of a registration form. The

server facilitates analyses using five popular algorithms and (currently) seven

different reference databases by accepting data in a user-friendly manner in a

wide range of formats, including those output by both commercial CD instruments

and synchrotron radiation-based circular dichroism beamlines, as well as those

produced by spectral processing software packages. It produces as output

calculated secondary structures, a goodness-of-fit parameter for the analyses,

and tabular and graphical displays of experimental, calculated and difference

spectra. The web pages associated with the server provide information on CD

spectroscopic methods and terms, literature references and aids for interpreting

the analysis results.

DOI: 10.1093/nar/gkh371

PMCID: PMC441509

PMID: 15215473 [Indexed for MEDLINE]

3283. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W665-7.

PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics

calculations.

Dolinsky TJ(1), Nielsen JE, McCammon JA, Baker NA.

Author information:

(1)Department of Biochemistry and Molecular Biophysics, Center for Computational

Biology, Washington University in St Louis, 700 S. Euclid Avenue, Campus Box

8036, St Louis, MO 63110, USA.

Continuum solvation models, such as Poisson-Boltzmann and Generalized Born

methods, have become increasingly popular tools for investigating the influence

of electrostatics on biomolecular structure, energetics and dynamics. However,

the use of such methods requires accurate and complete structural data as well as

force field parameters such as atomic charges and radii. Unfortunately, the

limiting step in continuum electrostatics calculations is often the addition of

missing atomic coordinates to molecular structures from the Protein Data Bank and

the assignment of parameters to biomolecular structures. To address this problem,

we have developed the PDB2PQR web service (http://agave.wustl.edu/pdb2pqr/). This

server automates many of the common tasks of preparing structures for continuum

electrostatics calculations, including adding a limited number of missing heavy

atoms to biomolecular structures, estimating titration states and protonating

biomolecules in a manner consistent with favorable hydrogen bonding, assigning

charge and radius parameters from a variety of force fields, and finally

generating 'PQR' output compatible with several popular computational biology

packages. This service is intended to facilitate the setup and execution of

electrostatics calculations for both experts and non-experts and thereby broaden

the accessibility to the biological community of continuum electrostatics

analyses of biomolecular systems.

DOI: 10.1093/nar/gkh381

PMCID: PMC441519

PMID: 15215472 [Indexed for MEDLINE]

3284. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W660-4.

PlasMapper: a web server for drawing and auto-annotating plasmid maps.

Dong X(1), Stothard P, Forsythe IJ, Wishart DS.

Author information:

(1)Department of Biological Sciences, University of Alberta, Edmonton, AB, T6G

2E8, Canada.

PlasMapper is a comprehensive web server that automatically generates and

annotates high-quality circular plasmid maps. Taking only the plasmid/vector DNA

sequence as input, PlasMapper uses sequence pattern matching and BLAST alignment

to automatically identify and label common promoters, terminators, cloning sites,

restriction sites, reporter genes, affinity tags, selectable marker genes,

replication origins and open reading frames. PlasMapper then presents the

identified features in textual form and as high-resolution, multicolored

graphical output. The appearance and contents of the output can be customized in

numerous ways using several supplied options. Further, PlasMapper images can be

rendered in both rasterized (PNG and JPG) and vector graphics (SVG) formats to

accommodate a variety of user needs or preferences. The images and textual output

are of sufficient quality that they may be used directly in publications or

presentations. The PlasMapper web server is freely accessible at

http://wishart.biology.ualberta.ca/PlasMapper.

DOI: 10.1093/nar/gkh410

PMCID: PMC441548

PMID: 15215471 [Indexed for MEDLINE]

3285. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W649-53.

ACMES: fast multiple-genome searches for short repeat sequences with concurrent

cross-species information retrieval.

Reneker J(1), Shyu CR, Zeng P, Polacco JC, Gassmann W.

Author information:

(1)Department of Computer Science, University of Missouri, Columbia, MO 65211,

USA.

We have developed a web server for the life sciences community to use to search

for short repeats of DNA sequence of length between 3 and 10,000 bases within

multiple species. This search employs a unique and fast hash function approach.

Our system also applies information retrieval algorithms to discover knowledge of

cross-species conservation of repeat sequences. Furthermore, we have incorporated

a part of the Gene Ontology database into our information retrieval algorithms to

broaden the coverage of the search. Our web server and tutorial can be found at

http://acmes.rnet.missouri.edu.

DOI: 10.1093/nar/gkh455

PMCID: PMC441593

PMID: 15215469 [Indexed for MEDLINE]

3286. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W645-8.

BSDD: Biomolecules Segment Display Device--a web-based interactive display tool.

Selvarani P(1), Shanthi V, Rajesh CK, Saravanan S, Sekar K.

Author information:

(1)Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India.

An interactive web-based display tool, Biomolecules Segment Display Device

(BSDD), has been developed to search for and visualize a user-defined motif or

fragment among the protein structures available in the Protein Data Bank (PDB).

In addition, the tool works for the structures available in a selected sub-set of

non-homologous protein structures (25% and 90% sequence identity). The graphics

package RASMOL has been incorporated as an interface to visualize the

three-dimensional structure of the user-defined motif. In addition, the software

can be used to extract the atomic coordinates of the required fragment and save

them to the client system. The atomic coordinates are updated every week from the

RCSB-PDB server, and hence the results produced by BSDD are up to date at any

given time. The software BSDD is available over the World Wide Web at

http://iris.physics.iisc.ernet.in/bsdd or http://144.16.71.2/bsdd.

DOI: 10.1093/nar/gkh420

PMCID: PMC441558

PMID: 15215468 [Indexed for MEDLINE]

3287. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W638-44.

ProMoST (Protein Modification Screening Tool): a web-based tool for mapping

protein modifications on two-dimensional gels.

Halligan BD(1), Ruotti V, Jin W, Laffoon S, Twigger SN, Dratz EA.

Author information:

(1)Bioinformatics Research Center, Medical College of Wisconsin, 8701 Watertown

Plank Road, Milwaukee, WI 53213, USA. Halligan@mcw.edu

ProMoST is a flexible web tool that calculates the effect of single or multiple

posttranslational modifications (PTMs) on protein isoelectric point (pI) and

molecular weight and displays the calculated patterns as two-dimensional (2D) gel

images. PTMs of proteins control many biological regulatory and signaling

mechanisms and 2D gel electrophoresis is able to resolve many PTM-induced

isoforms, such as those due to phosphorylation, acetylation, deamination,

alkylation, cysteine oxidation or tyrosine nitration. These modifications cause

changes in the pI of the protein by adding, removing or changing titratable

groups. Proteins differ widely in buffering capacity and pI and therefore the

same PTMs may give rise to quite different patterns of pI shifts in different

proteins. It is impossible by visual inspection of a pattern of spots on a gel to

determine which modifications are most likely to be present. The patterns of PTM

shifts for different proteins can be calculated and are often quite distinctive.

The theoretical gel images produced by ProMoST can be compared to the

experimental 2D gel results to implicate probable PTMs and focus efforts on more

detailed study of modified proteins. ProMoST has been implemented as cgi script

in Perl available on a WWW server at http://proteomics.mcw.edu/promost.

DOI: 10.1093/nar/gkh356

PMCID: PMC441494

PMID: 15215467 [Indexed for MEDLINE]

3288. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W634-7.

NLProt: extracting protein names and sequences from papers.

Mika S(1), Rost B.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 650 West 168th Street BB217, New York, NY 10032, USA.

mika@cubic.bioc.columbia.edu

Automatically extracting protein names from the literature and linking these

names to the associated entries in sequence databases is becoming increasingly

important for annotating biological databases. NLProt is a novel system that

combines dictionary- and rule-based filtering with several support vector

machines (SVMs) to tag protein names in PubMed abstracts. When considering

partially tagged names as errors, NLProt still reached a precision of 75% at a

recall of 76%. By many criteria our system outperformed other tagging methods

significantly; in particular, it proved very reliable even for novel names. Names

encountered particularly frequently in Drosophila, such as white, wing and

bizarre, constitute an obvious limitation of NLProt. Our method is available both

as an Internet server and as a program for download

(http://cubic.bioc.columbia.edu/services/NLProt/). Input can be PubMed/MEDLINE

identifiers, authors, titles and journals, as well as collections of abstracts,

or entire papers.

DOI: 10.1093/nar/gkh427

PMCID: PMC441565

PMID: 15215466 [Indexed for MEDLINE]

3289. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W624-7.

CSTminer: a web tool for the identification of coding and noncoding conserved

sequence tags through cross-species genome comparison.

Castrignanò T(1), Canali A, Grillo G, Liuni S, Mignone F, Pesole G.

Author information:

(1)Consorzio Interuniversitario per le Applicazioni di Supercalcolo per

Università e Ricerca, CASPUR, Rome, Italy.

The identification and characterization of genome tracts that are highly

conserved across species during evolution may contribute significantly to the

functional annotation of whole-genome sequences. Indeed, such sequences are

likely to correspond to known or unknown coding exons or regulatory motifs. Here,

we present a web server implementing a previously developed algorithm that, by

comparing user-submitted genome sequences, is able to identify statistically

significant conserved blocks and assess their coding or noncoding nature through

the measure of a coding potential score. The web tool, available at

http://www.caspur.it/CSTminer/, is dynamically interconnected with the Ensembl

genome resources and produces a graphical output showing a map of detected

conserved sequences and annotated gene features.

DOI: 10.1093/nar/gkh486

PMCID: PMC441624

PMID: 15215464 [Indexed for MEDLINE]

3290. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W620-3.

FAN: fingerprint analysis of nucleotide sequences.

Maudling N(1), Attwood TK.

Author information:

(1)School of Biological Sciences and Department of Computer Science, University

of Manchester, Manchester M13 9PT, UK. neil@bioinf.man.ac.uk

FAN is a server for fingerprint analysis of nucleotide sequences. The server

performs a search of submitted nucleotide sequences against the PRINTS database.

Searches are performed directly against fingerprints using a codon position

specific score matrix (PSSM) approach. The advantages of this approach are

increased specificity for coding sequence (CDS) over non-CDS, and increased

tolerance to base-substituting and frameshifting sequence errors. Furthermore,

there is no need for prior translation of the nucleotide sequences. A web-based

interface to the software is available at

http://bioinf.man.ac.uk/cgi-bin/neil/ntfront.pl.

DOI: 10.1093/nar/gkh457

PMCID: PMC441595

PMID: 15215463 [Indexed for MEDLINE]

3291. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W610-4.

ElNemo: a normal mode web server for protein movement analysis and the generation

of templates for molecular replacement.

Suhre K(1), Sanejouand YH.

Author information:

(1)Information Génomique & Structurale (UPR CNRS 2589), 31, chemin Joseph

Aiguier, 13402 Marseille Cedex 20, France. karsten.suhre@igs.cnrs-mrs.fr

Normal mode analysis (NMA) is a powerful tool for predicting the possible

movements of a given macromolecule. It has been shown recently that half of the

known protein movements can be modelled by using at most two low-frequency normal

modes. Applications of NMA cover wide areas of structural biology, such as the

study of protein conformational changes upon ligand binding, membrane channel

opening and closure, potential movements of the ribosome, and viral capsid

maturation. Another, newly emerging field of NMA is related to protein structure

determination by X-ray crystallography, where normal mode perturbed models are

used as templates for diffraction data phasing through molecular replacement

(MR). Here we present ElNémo, a web interface to the Elastic Network Model that

provides a fast and simple tool to compute, visualize and analyse low-frequency

normal modes of large macro-molecules and to generate a large number of different

starting models for use in MR. Due to the 'rotation-translation-block' (RTB)

approximation implemented in ElNémo, there is virtually no upper limit to the

size of the proteins that can be treated. Upon input of a protein structure in

Protein Data Bank (PDB) format, ElNémo computes its 100 lowest-frequency modes

and produces a comprehensive set of descriptive parameters and visualizations,

such as the degree of collectivity of movement, residue mean square

displacements, distance fluctuation maps, and the correlation between observed

and normal-mode-derived atomic displacement parameters (B-factors). Any number of

normal mode perturbed models for MR can be generated for download. If two

conformations of the same (or a homologous) protein are available, ElNémo

identifies the normal modes that contribute most to the corresponding protein

movement. The web server can be freely accessed at

http://igs-server.cnrs-mrs.fr/elnemo/index.html.

DOI: 10.1093/nar/gkh368

PMCID: PMC441506

PMID: 15215461 [Indexed for MEDLINE]

3292. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W606-9.

CaspR: a web server for automated molecular replacement using homology modelling.

Claude JB(1), Suhre K, Notredame C, Claverie JM, Abergel C.

Author information:

(1)Information Génomique & Structurale (UPR CNRS 2589), Institut de Biologie

Structurale et Microbiologie, 31, chemin Joseph Aiguier, 13402 Marseille Cedex

20, France.

Molecular replacement (MR) is the method of choice for X-ray crystallography

structure determination when structural homologues are available in the Protein

Data Bank (PDB). Although the success rate of MR decreases sharply when the

sequence similarity between template and target proteins drops below 35%

identical residues, it has been found that screening for MR solutions with a

large number of different homology models may still produce a suitable solution

where the original template failed. Here we present the web tool CaspR,

implementing such a strategy in an automated manner. On input of experimental

diffraction data, of the corresponding target sequence and of one or several

potential templates, CaspR executes an optimized molecular replacement procedure

using a combination of well-established stand-alone software tools. The protocol

of model building and screening begins with the generation of multiple

structure-sequence alignments produced with T-COFFEE, followed by homology model

building using MODELLER, molecular replacement with AMoRe and model refinement

based on CNS. As a result, CaspR provides a progress report in the form of

hierarchically organized summary sheets that describe the different stages of the

computation with an increasing level of detail. For the 10 highest-scoring

potential solutions, pre-refined structures are made available for download in

PDB format. Results already obtained with CaspR and reported on the web server

suggest that such a strategy significantly increases the fraction of protein

structures which may be solved by MR. Moreover, even in situations where standard

MR yields a solution, pre-refined homology models produced by CaspR significantly

reduce the time-consuming refinement process. We expect this automated procedure

to have a significant impact on the throughput of large-scale structural genomics

projects. CaspR is freely available at http://igs-server.cnrs-mrs.fr/Caspr/.

DOI: 10.1093/nar/gkh400

PMCID: PMC441538

PMID: 15215460 [Indexed for MEDLINE]

3293. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W602-5.

iMOT: an interactive package for the selection of spatially interacting motifs.

Bhaduri A(1), Pugalenthi G, Gupta N, Sowdhamini R.

Author information:

(1)National Centre for Biological Sciences, Tata Institute of Fundamental

Research, UAS-GKVK Campus, Bellary Road, Bangalore 560 065, Karnataka, India.

Functional selection and three-dimensional structural constraints of proteins

relate to the retention of significant sequence similarity between proteins of

similar fold and function despite poor overall sequence identity and evolutionary

pressures. We report the availability of 'iMOT' (interacting MOTif) server, an

interactive package for the automatic identification of spatially interacting

motifs among distantly related proteins sharing similar folds and possessing

common ancestral lineage. Spatial interactions between conserved stretches of a

protein are evaluated by calculations of pseudo-potentials that describe the

strength of interactions. Such an evaluation permits the automatic identification

of highly interacting conserved regions of a protein. Interacting motifs have

been shown to be useful in searching for distant homologues and establishing

remote homologies among the largely unassigned sequences in genome databases.

Information on such motifs should also be of value in protein folding, modelling

and engineering experiments. The iMOT server can be accessed from

http://www.ncbs.res.in/~faculty/mini/imot/iMOTserver.html. Supplementary Material

can be accessed from:

http://www.ncbs.res.in/~faculty/mini/imot/supplementary.html.

DOI: 10.1093/nar/gkh375

PMCID: PMC441513

PMID: 15215459 [Indexed for MEDLINE]

3294. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W590-4.

SuperPose: a simple server for sophisticated structural superposition.

Maiti R(1), Van Domselaar GH, Zhang H, Wishart DS.

Author information:

(1)Departments of Biological Sciences and Computing Science, University of

Alberta, Edmonton, AB, T6G 2E8, Canada.

The SuperPose web server rapidly and robustly calculates both pairwise and

multiple protein structure superpositions using a modified quaternion eigenvalue

approach. SuperPose generates sequence alignments, structure alignments, PDB

(Protein Data Bank) coordinates and RMSD statistics, as well as difference

distance plots and images (both static and interactive) of the superimposed

molecules. SuperPose employs a simple interface that requires only PDB files or

accession numbers as input. All other superposition decisions are made by the

program. SuperPose is uniquely able to superimpose structures that differ

substantially in sequence, size or shape. It is also capable of handling a much

larger range of superposition queries and situations than many standalone

programs and yields results that are intuitively more in agreement with known

biological or structural data. The SuperPose web server is freely accessible at

http://wishart.biology.ualberta.ca/SuperPose/.

DOI: 10.1093/nar/gkh477

PMCID: PMC441615

PMID: 15215457 [Indexed for MEDLINE]

3295. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W586-9.

COLORADO3D, a web server for the visual analysis of protein structures.

Sasin JM(1), Bujnicki JM.

Author information:

(1)Laboratory of Bioinformatics and Protein Engineering, International Institute

of Molecular and Cell Biology, Trojdena 4, Warsaw, Poland.

COLORADO3D is a World Wide Web server for the visual presentation of

three-dimensional (3D) protein structures. COLORADO3D indicates the presence of

potential errors (detected by ANOLEA, PROSAII, PROVE or VERIFY3D), identifies

buried residues and depicts sequence conservations. As input, the server takes a

file of Protein Data Bank (PDB) coordinates and, optionally, a multiple sequence

alignment. As output, the server returns a PDB-formatted file, replacing the

B-factor column with values of the chosen parameter (structure quality, residue

burial or conservation). Thus, the coordinates of the analyzed protein 'colored'

by COLORADO3D can be conveniently displayed with structure viewers such as RASMOL

in order to visualize the 3D clusters of regions with common features, which may

not necessarily be adjacent to each other at the amino acid sequence level. In

particular, COLORADO3D may serve as a tool to judge a structure's quality at

various stages of the modeling and refinement (during both experimental structure

determination and homology modeling). The GeneSilico group used COLORADO3D in the

fifth Critical Assessment of Techniques for Protein Structure Prediction (CASP5)

to successfully identify well-folded parts of preliminary homology models and to

guide the refinement of misthreaded protein sequences. COLORADO3D is freely

available for academic use at http://asia.genesilico.pl/colorado3d/.

DOI: 10.1093/nar/gkh440

PMCID: PMC441578

PMID: 15215456 [Indexed for MEDLINE]

3296. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W582-5.

FATCAT: a web server for flexible structure comparison and structure similarity

searching.

Ye Y(1), Godzik A.

Author information:

(1)The Burnham Institute, 10901 N. Torrey Pines Road, La Jolla, CA 92037, USA.

yye@burnham.org

Protein structure comparison, an important problem in structural biology, has two

main applications: (i) comparing two protein structures in order to identify the

similarities and differences between them, and (ii) searching for structures

similar to a query structure. Many web-based resources for both applications are

available, but all are based on rigid structural alignment algorithms. FATCAT

server implements the recently developed flexible protein structure comparison

algorithm FATCAT, which automatically identifies hinges and internal

rearrangements in two protein structures. The server provides access to two

algorithms: FATCAT-pairwise for pairwise flexible structure comparison and

FATCAT-search for database searching for structurally similar proteins. Given two

protein structures [in the Protein Data Bank (PDB) format], FATCAT-pairwise

reports their structural alignment and the corresponding statistical significance

of the similarity measured as a P-value. Users can view the superposition of the

structures online in web browsers that support the Chime plug-in, or download the

superimposed structures in PDB format. In FATCAT-search, users provide one query

structure and the server returns a list of protein structures that are similar to

the query, ordered by the P-values. In addition, FATCAT server can report the

conformational changes of the query structure as compared to other proteins in

the structure database. FATCAT server is available at http://fatcat.burnham.org.

DOI: 10.1093/nar/gkh430

PMCID: PMC441568

PMID: 15215455 [Indexed for MEDLINE]

3297. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W572-5.

ProteinDBS: a real-time retrieval system for protein structure comparison.

Shyu CR(1), Chi PH, Scott G, Xu D.

Author information:

(1)Medical and Biological Digital Library Research Lab, 322 Engineering Building

North, Department of Computer Science, University of Missouri, Columbia, MO

65211-2060, USA. shyuc@missouri.edu

We have developed a web server (ProteinDBS) for the life science community to

search for similar protein tertiary structures in real time. This system applies

computer visualization techniques to extract the predominant visual patterns

encoded in two-dimensional distance matrices generated from the three-dimensional

coordinates of protein chains. When meaningful contents, represented in a

multi-dimensional feature space, have been extracted from distance matrices, an

advanced indexing structure, Entropy Balanced Statistical (EBS) k-d tree, is

utilized to index the data. Our system is able to return search results in ranked

order from a database with 46 075 chains in seconds, exhibiting a reasonably high

degree of precision. To our knowledge, this is the first real-time search engine

for protein structure comparison. ProteinDBS provides two types of query method:

query by Protein Data Bank protein chain ID and by new structures uploaded by

users. The system is hosted at http://ProteinDBS.rnet.missouri.edu.

DOI: 10.1093/nar/gkh436

PMCID: PMC441574

PMID: 15215453 [Indexed for MEDLINE]

3298. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W569-71.

CHOP: parsing proteins into structural domains.

Liu J(1), Rost B.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 650 West 168th Street BB217, New York, NY 10032, USA.

Sequence-based domain assignment is one of the most important and challenging

problems in structural biology. We have developed a method, CHOP, that chops

proteins into domain-like fragments. The basic idea is to cut proteins from

entirely sequenced organisms beginning from very reliable experimental

information (Protein Data Bank), proceeding to expert annotations of domain-like

regions (Pfam-A) and completing through cuts based on termini of native protein

ends. The CHOP server takes protein sequences as input and returns the

dissections supported by homology transfer. CHOP results are precompiled for many

entirely sequenced proteomes. The service is available at

http://www.rostlab.org/services/CHOP/.

DOI: 10.1093/nar/gkh481

PMCID: PMC441619

PMID: 15215452 [Indexed for MEDLINE]

3299. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W562-5.

LINKER: a web server to generate peptide sequences with extended conformation.

Xue F(1), Gu Z, Feng JA.

Author information:

(1)Department of Chemistry, Temple University, 1901 N 13th Street, Philadelphia,

PA 19122, USA.

LINKER was developed as an online server to assist biomedical researchers to

design linker sequences for constructing functional fusion proteins. The program

automatically generates a set of peptide sequences that are known to adopt

extended conformations as determined by X-ray crystallography and NMR. In

addition to the desired linker sequence length, the web interface provides a

number of optional input parameters so that the users may enhance sequence

selection based on the requirements of specific applications. The output of

LINKER includes a list of peptide sequences with specified length and sequence

characteristics. A graphical subroutine was implemented to highlight the chemical

features of every linker sequence by hydrophobicity plots. LINKER can be accessed

at http://astro.temple.edu/~feng/Servers/BioinformaticServers.htm.

DOI: 10.1093/nar/gkh422

PMCID: PMC441560

PMID: 15215450 [Indexed for MEDLINE]

3300. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W559-61.

PepBuild: a web server for building structure data of peptides/proteins.

Singh B(1).

Author information:

(1)Bioinformatics, Institute of Microbial Technology, Sector 39A, Chandigarh

160036, India. bvs@imtech.res.in

PepBuild, a web server, will aid in designing and building a capped or uncapped

peptide/protein with known secondary and tertiary structure. The user can build a

peptide/protein by choosing the required amino acid residue with regular

secondary structure. The torsional angles can be supplied by the user, if

desired. The server also allows the user to add relevant protecting groups at the

N- and/or C-terminal of the peptide. The amino acid side chains of the designed

peptide are optimized using rotameric libraries. Finally, the server provides the

option of displaying the result or downloading the complete file in PDB (Protein

Data Bank) format. This PDB file can later be used as an input for various

molecular simulation programs or for graphical display. The web server is

available at http://www.imtech.res.in/bvs/pepbuild/.

DOI: 10.1093/nar/gkh472

PMCID: PMC441610

PMID: 15215449 [Indexed for MEDLINE]

3301. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W555-8.

pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino

acid residues on proteins.

Binkowski TA(1), Freeman P, Liang J.

Author information:

(1)Department of Bioengineering, The University of Illinois at Chicago, Chicago,

IL 60607-7052, USA.

Detecting similar protein surfaces provides an important route for discovering

unrecognized or novel functional relationship between proteins. The web server

pvSOAR (pocket and void Surfaces Of Amino acid Residues) provides an online

resource to identify similar protein surface regions. pvSOAR can take a structure

either uploaded by a user or obtained from the Protein Data Bank, and identifies

similar surface patterns based on geometrically defined pockets and voids. It

provides several search modes to compare protein surfaces by similarity in local

sequence, local shape and local orientation. Statistically significant search

results are reported for visualization and interactive exploration. pvSOAR can be

used to predict biological functions of proteins with known three-dimensional

structures but unknown biological roles. It can also be used to study functional

relationship between proteins and for exploration of the evolutionary origins of

structural elements important for protein function. We present an example using

pvSOAR to explore the biological roles of a protein whose structure was solved by

the structural genomics project. The pvSOAR web server is available at

http://pvsoar.bioengr.uic.edu/.

DOI: 10.1093/nar/gkh390

PMCID: PMC441528

PMID: 15215448 [Indexed for MEDLINE]

3302. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W542-4.

SledgeHMMER: a web server for batch searching the Pfam database.

Chukkapalli G(1), Guda C, Subramaniam S.

Author information:

(1)San Diego Supercomputer Center, University of California, San Diego, 9500

Gilman Drive, La Jolla, CA 92093-0505, USA.

The SledgeHMMER web server is intended for genome-scale searching of the Pfam

database without having to install this database and the HMMER software locally.

The server implements a parallelized version of hmmpfam, the program used for

searching the Pfam HMM database. Pfam search results have been calculated for the

entire Swiss-Prot and TrEmbl database sequences (approximately 1.2 million) on

256 processors of IA64-based teragrid machines. The Pfam database can be searched

in local, glocal or merged mode, using either gathering or E-value thresholds.

Query sequences are first matched against the pre-calculated entries to retrieve

results, and those without matches are processed through a new search process.

Results are emailed in a space-delimited tabular format upon completion of the

search. While most other Pfam-searching web servers set a limit of one sequence

per query, this server processes batch sequences with no limit on the number of

input sequences. The web server and downloadable data are accessible from

http://SledgeHmmer.sdsc.edu.

DOI: 10.1093/nar/gkh395

PMCID: PMC441533

PMID: 15215445 [Indexed for MEDLINE]

3303. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W536-41.

FoldMiner and LOCK 2: protein structure comparison and motif discovery on the

web.

Shapiro J(1), Brutlag D.

Author information:

(1)Biophysics Program, Stanford University School of Medicine, Stanford, CA

94305-5307, USA.

The FoldMiner web server (http://foldminer.stanford.edu/) provides remote access

to methods for protein structure alignment and unsupervised motif discovery.

FoldMiner is unique among such algorithms in that it improves both the motif

definition and the sensitivity of a structural similarity search by combining the

search and motif discovery methods and using information from each process to

enhance the other. In a typical run, a query structure is aligned to all

structures in one of several databases of single domain targets in order to

identify its structural neighbors and to discover a motif that is the basis for

the similarity among the query and statistically significant targets. This

process is fully automated, but options for manual refinement of the results are

available as well. The server uses the Chime plugin and customized controls to

allow for visualization of the motif and of structural superpositions. In

addition, we provide an interface to the LOCK 2 algorithm for rapid alignments of

a query structure to smaller numbers of user-specified targets.

DOI: 10.1093/nar/gkh389

PMCID: PMC441527

PMID: 15215444 [Indexed for MEDLINE]

3304. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W532-5.

Wurst: a protein threading server with a structural scoring function, sequence

profiles and optimized substitution matrices.

Torda AE(1), Procter JB, Huber T.

Author information:

(1)University of Hamburg, Zentrum für Bioinformatik, Bundesstrasse 43, D-20146

Hamburg, Germany.

Wurst is a protein threading program with an emphasis on high quality sequence to

structure alignments (http://www.zbh.uni-hamburg.de/wurst). Submitted sequences

are aligned to each of about 3000 templates with a conventional dynamic

programming algorithm, but using a score function with sophisticated structure

and sequence terms. The structure terms are a log-odds probability of sequence to

structure fragment compatibility, obtained from a Bayesian classification

procedure. A simplex optimization was used to optimize the sequence-based terms

for the goal of alignment and model quality and to balance the sequence and

structural contributions against each other. Both sequence and structural terms

operate with sequence profiles.

DOI: 10.1093/nar/gkh357

PMCID: PMC441495

PMID: 15215443 [Indexed for MEDLINE]

3305. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W526-31.

Protein structure prediction and analysis using the Robetta server.

Kim DE(1), Chivian D, Baker D.

Author information:

(1)Structural Genomics of Pathogenic Protozoa, Department of Biochemistry,

University of Washington, Seattle WA 98195, USA.

The Robetta server (http://robetta.bakerlab.org) provides automated tools for

protein structure prediction and analysis. For structure prediction, sequences

submitted to the server are parsed into putative domains and structural models

are generated using either comparative modeling or de novo structure prediction

methods. If a confident match to a protein of known structure is found using

BLAST, PSI-BLAST, FFAS03 or 3D-Jury, it is used as a template for comparative

modeling. If no match is found, structure predictions are made using the de novo

Rosetta fragment insertion method. Experimental nuclear magnetic resonance (NMR)

constraints data can also be submitted with a query sequence for RosettaNMR de

novo structure determination. Other current capabilities include the prediction

of the effects of mutations on protein-protein interactions using computational

interface alanine scanning. The Rosetta protein design and protein-protein

docking methodologies will soon be available through the server as well.

DOI: 10.1093/nar/gkh468

PMCID: PMC441606

PMID: 15215442 [Indexed for MEDLINE]

3306. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W517-21.

LOCnet and LOCtarget: sub-cellular localization for structural genomics targets.

Nair R(1), Rost B.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 650 West 168th Street BB217, New York, NY 10032, USA.

nair@maple.bioc.columbia.edu

LOCtarget is a web server and database that predicts and annotates sub-cellular

localization for structural genomics targets; LOCnet is one of the methods used

in LOCtarget that can predict sub-cellular localization for all eukaryotic and

prokaryotic proteins. Targets are taken from the central registration database

for structural genomics, namely, TargetDB. LOCtarget predicts localization

through a combination of four different methods: known nuclear localization

signals (PredictNLS), homology-based transfer of experimental annotations

(LOChom), inference through automatic text analysis of SWISS-PROT keywords

(LOCkey) and de novo prediction through a system of neural networks (LOCnet).

Additionally, we report predictions from SignalP. The final prediction is based

on the method with the highest confidence. The web server can be used to predict

sub-cellular localization of proteins from their amino acid sequence. The

LOCtarget database currently contains localization predictions for all eukaryotic

proteins from TargetDB and is updated every week. The server is available at

http://www.rostlab.org/services/LOCtarget/.

DOI: 10.1093/nar/gkh441

PMCID: PMC441579

PMID: 15215440 [Indexed for MEDLINE]

3307. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W512-6.

iMolTalk: an interactive, internet-based protein structure analysis server.

Diemand AV(1), Scheib H.

Author information:

(1)University of Lausanne and Swiss Institute of Bioinformatics, Chemin des

Boveresses 155, 1066 Epalinges s/Lausanne, Switzerland.

alexander.diemand@isb-sib.ch

iMolTalk (http://i.moltalk.org) is a new and interactive web server for protein

structure analysis. It addresses the need to identify and highlight biochemically

important regions in protein structures. As input, the server requires only the

four-digit Protein Data Bank (PDB) identifier, of an experimentally determined

structure or a structure file in PDB format stemming e.g. from comparative

modelling. iMolTalk offers a wide range of implemented tools (i) to extract

general information from PDB files, such as generic header information or the

sequence derived from three-dimensional co-ordinates; (ii) to map corresponding

residues from sequence to structure; (iii) to search for contacts of residues

(amino or nucleic acids) or heterogeneous groups to the protein, present

cofactors and substrates; and (iv) to identify protein-protein interfaces between

chains in a structure. The server provides results as user-friendly

two-dimensional graphical representations and in textual format, ideal for

further processing. At any time during the analysis, the user can choose, for the

following step, from the set of implemented tools or submit his/her own script to

the server to extend the functionality of iMolTalk.

DOI: 10.1093/nar/gkh403

PMCID: PMC441541

PMID: 15215439 [Indexed for MEDLINE]

3308. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W508-11.

SCit: web tools for protein side chain conformation analysis.

Gautier R(1), Camproux AC, Tufféry P.

Author information:

(1)Equipe de Bioinformatique Génomique et Moléculaire, INSERM E346, Université

Paris 7, case 7113, 2, place Jussieu, 75251 Paris cedex 05, France.

SCit is a web server providing services for protein side chain conformation

analysis and side chain positioning. Specific services use the dependence of the

side chain conformations on the local backbone conformation, which is described

using a structural alphabet that describes the conformation of fragments of

four-residue length in a limited library of structural prototypes. Based on this

concept, SCit uses sets of rotameric conformations dependent on the local

backbone conformation of each protein for side chain positioning and the

identification of side chains with unlikely conformations. The SCit web server is

accessible at http://bioserv.rpbs.jussieu.fr/SCit.

DOI: 10.1093/nar/gkh388

PMCID: PMC441526

PMID: 15215438 [Indexed for MEDLINE]

3309. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W503-7.

BioInfo3D: a suite of tools for structural bioinformatics.

Shatsky M(1), Dror O, Schneidman-Duhovny D, Nussinov R, Wolfson HJ.

Author information:

(1)School of Computer Science, Raymond and Beverly Sackler Faculty of Exact

Sciences,Tel Aviv University, Tel Aviv 69978, Israel.

Here, we describe BioInfo3D, a suite of freely available web services for protein

structural analysis. The FlexProt method performs flexible structural alignment

of protein molecules. FlexProt simultaneously detects the hinge regions and

aligns the rigid subparts of the molecules. It does not require an a priori

knowledge of the flexible hinge regions. MultiProt and MASS perform simultaneous

comparison of multiple protein structures. PatchDock performs prediction of

protein-protein and protein-small molecule interactions. The input to all

services is either protein PDB codes or protein structures uploaded to the

server. All the services are available at http://bioinfo3d.cs.tau.ac.il.

DOI: 10.1093/nar/gkh413

PMCID: PMC441551

PMID: 15215437 [Indexed for MEDLINE]

3310. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W500-2.

STRIDE: a web server for secondary structure assignment from known atomic

coordinates of proteins.

Heinig M(1), Frishman D.

Author information:

(1)Department of Genome Oriented Bioinformatics, Technical University of Munich,

Wissenschaftszentrum Weihenstephan, 85354 Freising, Germany.

STRIDE is a software tool for secondary structure assignment from atomic

resolution protein structures. It implements a knowledge-based algorithm that

makes combined use of hydrogen bond energy and statistically derived backbone

torsional angle information and is optimized to return resulting assignments in

maximal agreement with crystallographers' designations. The STRIDE web server

provides access to this tool and allows visualization of the secondary structure,

as well as contact and Ramachandran maps for any file uploaded by the user with

atomic coordinates in the Protein Data Bank (PDB) format. A searchable database

of STRIDE assignments for the latest PDB release is also provided. The STRIDE

server is accessible from http://webclu.bio.wzw.tum.de/stride/.

DOI: 10.1093/nar/gkh429

PMCID: PMC441567

PMID: 15215436 [Indexed for MEDLINE]

3311. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W457-9.

ArrayPipe: a flexible processing pipeline for microarray data.

Hokamp K(1), Roche FM, Acab M, Rousseau ME, Kuo B, Goode D, Aeschliman D, Bryan

J, Babiuk LA, Hancock RE, Brinkman FS.

Author information:

(1)Department of Molecular Biology and Biochemistry, Simon Fraser University,

Burnaby, BC, Canada.

A number of microarray analysis software packages exist already; however, none

combines the user-friendly features of a web-based interface with potential

ability to analyse multiple arrays at once using flexible analysis steps. The

ArrayPipe web server (freely available at www.pathogenomics.ca/arraypipe) allows

the automated application of complex analyses to microarray data which can range

from single slides to large data sets including replicates and dye-swaps. It

handles output from most commonly used quantification software packages for

dual-labelled arrays. Application features range from quality assessment of

slides through various data visualizations to multi-step analyses including

normalization, detection of differentially expressed genes, andcomparison and

highlighting of gene lists. A highly customizable action set-up facilitates

unrestricted arrangement of functions, which can be stored as action profiles. A

unique combination of web-based and command-line functionality enables

comfortable configuration of processes that can be repeatedly applied to large

data sets in high throughput. The output consists of reports formatted as

standard web pages and tab-delimited lists of calculated values that can be

inserted into other analysis programs. Additional features, such as web-based

spreadsheet functionality, auto-parallelization and password protection make this

a powerful tool in microarray research for individuals and large groups alike.

DOI: 10.1093/nar/gkh446

PMCID: PMC441584

PMID: 15215429 [Indexed for MEDLINE]

3312. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W441-4.

KARMA: a web server application for comparing and annotating heterogeneous

microarray platforms.

Cheung KH(1), Hager J, Pan D, Srivastava R, Mane S, Li Y, Miller P, Williams KR.

Author information:

(1)Center for Medical Informatics, Department of Anesthesiology, Yale University,

New Haven, CT, USA. kei.cheung@yale.edu

We have developed a universal web server application (KARMA) that allows

comparison and annotation of user-defined pairs of microarray platforms based on

diverse types of genome annotation data (across different species) collected from

multiple sources. The application is an effective tool for diverse microarray

platforms, including arrays that are provided by (i) the Keck Microarray Resource

at Yale, (ii) commercially available Affymetrix GeneChips and spotted arrays and

(iii) custom arrays made by individual academics. The tool provides a web

interface that allows users to input pairs of test files that represent diverse

array platforms for either single or multiple species. The program dynamically

identifies analogous DNA fragments spotted or synthesized on multiple microarray

platforms based on the following types of information: (i) NCBI-Unigene

identifiers, if the platforms being compared are within the same species or (ii)

NCBI-Homologene data, if they are cross-species. The single-species comparison is

implemented based on set operations: intersection, union and difference. Other

forms of retrievable annotation data, including LocusLink, SwissProt and Gene

Ontology (GO), are collected from multiple remote sites and stored in an

integrated fashion using an Oracle database. The KARMA database, which is updated

periodically, is available on line at the following URL:

http://ymd.med.yale.edu/karma/cgi-bin/karma.pl.

DOI: 10.1093/nar/gkh397

PMCID: PMC441535

PMID: 15215426 [Indexed for MEDLINE]

3313. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W429-34.

The Iccare web server: an attempt to merge sequence and mapping information for

plant and animal species.

Muller C(1), Denis M, Gentzbittel L, Faraut T.

Author information:

(1)INP-ENSAT, Laboratoire de biotechnologies et d'amélioration des plantes,

Castanet Tolosan 31326, France.

The Iccare web server, http://genopole.toulouse.inra.fr/bioinfo/Iccare, provides

a simple yet efficient tool for crude EST (expressed sequence tag) annotation

specifically dedicated to comparative mapping approaches. Iccare uses all the EST

and mRNA sequences from public databases for an organism of interest (query

species) and compares them to all the transcripts of one reference organism (Homo

sapiens or Arabidopsis thaliana). The results are displayed according to the

location of the genes on the chromosomes of the reference organism. Gene

structure information and sequence similarities are combined in a graphical

representation in order to pinpoint the nature of the transcript query sequence.

The user can subsequently design primers or probes for the purpose of physical or

genetic mapping. In addition to the query organisms already available in Iccare,

users can perform a tailor-made search with their own sequences against the

animal or plant reference organism genes.

DOI: 10.1093/nar/gkh460

PMCID: PMC441598

PMID: 15215424 [Indexed for MEDLINE]

3314. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W420-3.

MotifViz: an analysis and visualization tool for motif discovery.

Fu Y(1), Frith MC, Haverty PM, Weng Z.

Author information:

(1)Bioinformatics Program, Boston University, Boston, MA 02215, USA.

Detecting overrepresented known transcription factor binding motifs in a set of

promoter sequences of co-regulated genes has become an important approach to

deciphering transcriptional regulatory mechanisms. In this paper, we present an

interactive web server, MotifViz, for three motif discovery programs, Clover,

Rover and Motifish, covering most available flavors of algorithms for achieving

this goal. For comparison, we have also implemented the simple motif-matching

program Possum. MotifViz provides uniform and intuitive input and output formats

for all four programs. It can be accessed at http://biowulf.bu.edu/MotifViz.

DOI: 10.1093/nar/gkh426

PMCID: PMC441564

PMID: 15215422 [Indexed for MEDLINE]

3315. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W414-9.

ESLpred: SVM-based method for subcellular localization of eukaryotic proteins

using dipeptide composition and PSI-BLAST.

Bhasin M(1), Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39A,

Chandigarh, India.

Automated prediction of subcellular localization of proteins is an important step

in the functional annotation of genomes. The existing subcellular localization

prediction methods are based on either amino acid composition or N-terminal

characteristics of the proteins. In this paper, support vector machine (SVM) has

been used to predict the subcellular location of eukaryotic proteins from their

different features such as amino acid composition, dipeptide composition and

physico-chemical properties. The SVM module based on dipeptide composition

performed better than the SVM modules based on amino acid composition or

physico-chemical properties. In addition, PSI-BLAST was also used to search the

query sequence against the dataset of proteins (experimentally annotated

proteins) to predict its subcellular location. In order to improve the prediction

accuracy, we developed a hybrid module using all features of a protein, which

consisted of an input vector of 458 dimensions (400 dipeptide compositions, 33

properties, 20 amino acid compositions of the protein and 5 from PSI-BLAST

output). Using this hybrid approach, the prediction accuracies of nuclear,

cytoplasmic, mitochondrial and extracellular proteins reached 95.3, 85.2, 68.2

and 88.9%, respectively. The overall prediction accuracy of SVM modules based on

amino acid composition, physico-chemical properties, dipeptide composition and

the hybrid approach was 78.1, 77.8, 82.9 and 88.0%, respectively. The accuracy of

all the modules was evaluated using a 5-fold cross-validation technique.

Assigning a reliability index (reliability index > or =3), 73.5% of prediction

can be made with an accuracy of 96.4%. Based on the above approach, an online web

server ESLpred was developed, which is available at

http://www.imtech.res.in/raghava/eslpred/.

DOI: 10.1093/nar/gkh350

PMCID: PMC441488

PMID: 15215421 [Indexed for MEDLINE]

3316. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W400-4.

PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane

proteins.

Bagos PG(1), Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ.

Author information:

(1)Department of Cell Biology and Biophysics, Faculty of Biology, University of

Athens, Panepistimiopolis, Athens 15701, Greece. pbagos@biol.uoa.gr

The beta-barrel outer membrane proteins constitute one of the two known

structural classes of membrane proteins. Whereas there are several different

web-based predictors for alpha-helical membrane proteins, currently there is no

freely available prediction method for beta-barrel membrane proteins, at least

with an acceptable level of accuracy. We present here a web server (PRED-TMBB,

http://bioinformatics.biol.uoa.gr/PRED-TMBB) which is capable of predicting the

transmembrane strands and the topology of beta-barrel outer membrane proteins of

Gram-negative bacteria. The method is based on a Hidden Markov Model, trained

according to the Conditional Maximum Likelihood criterion. The model was

retrained and the training set now includes 16 non-homologous outer membrane

proteins with structures known at atomic resolution. The user may submit one

sequence at a time and has the option of choosing between three different

decoding methods. The server reports the predicted topology of a given protein, a

score indicating the probability of the protein being an outer membrane

beta-barrel protein, posterior probabilities for the transmembrane strand

prediction and a graphical representation of the assumed position of the

transmembrane strands with respect to the lipid bilayer.

DOI: 10.1093/nar/gkh417

PMCID: PMC441555

PMID: 15215419 [Indexed for MEDLINE]

3317. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W390-3.

ConPred II: a consensus prediction method for obtaining transmembrane topology

models with high reliability.

Arai M(1), Mitsuke H, Ikeda M, Xia JX, Kikuchi T, Satake M, Shimizu T.

Author information:

(1)Department of Electronic and Information System Engineering, Faculty of

Science and Technology, Hirosaki University, Hirosaki 036-8561, Japan.

ConPred II (http://bioinfo.si.hirosaki-u.ac.jp/~ConPred2/) is a server for the

prediction of transmembrane (TM) topology [i.e. the number of TM segments (TMSs),

TMS positions and N-tail location] based on a consensus approach by combining the

results of several proposed methods. The ConPred II system is constructed from

ConPred\_elite and ConPred\_all (previously named ConPred), proposed earlier by our

group. The prediction accuracy of ConPred\_elite is almost 100%, which is achieved

by sacrificing the prediction coverage (20-30%). ConPred\_all predicts TM

topologies for all the input sequences with accuracies improved by up to 11% over

individual proposed methods. In the ConPred II system, the TM topology prediction

of input TM protein sequences is executed following a two-step process: (i) input

sequences are first run through the ConPred\_elite program; (ii) sequences for

which ConPred\_elite does not give the TM topology are delivered to the

ConPred\_all program for TM topology prediction. Users can get access to the

ConPred II system automatically by submitting sequences to the server. The

ConPred II server will return the predicted TM topology models and graphical

representations of their contents (hydropathy plots, helical wheel diagrams of

predicted TMSs and snake-like diagrams).

DOI: 10.1093/nar/gkh380

PMCID: PMC441518

PMID: 15215417 [Indexed for MEDLINE]

3318. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W383-9.

GPCRpred: an SVM-based method for prediction of families and subfamilies of

G-protein coupled receptors.

Bhasin M(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology Sector 39-A, Chandigarh, 160036, India.

G-protein coupled receptors (GPCRs) belong to one of the largest superfamilies of

membrane proteins and are important targets for drug design. In this study, a

support vector machine (SVM)-based method, GPCRpred, has been developed for

predicting families and subfamilies of GPCRs from the dipeptide composition of

proteins. The dataset used in this study for training and testing was obtained

from http://www.soe.ucsc.edu/research/compbio/gpcr/. The method classified GPCRs

and non-GPCRs with an accuracy of 99.5% when evaluated using 5-fold

cross-validation. The method is further able to predict five major classes or

families of GPCRs with an overall Matthew's correlation coefficient (MCC) and

accuracy of 0.81 and 97.5% respectively. In recognizing the subfamilies of the

rhodopsin-like family, the method achieved an average MCC and accuracy of 0.97

and 97.3% respectively. The method achieved overall accuracy of 91.3% and 96.4%

at family and subfamily level respectively when evaluated on an independent/blind

dataset of 650 GPCRs. A server for recognition and classification of GPCRs based

on multiclass SVMs has been set up at http://www.imtech.res.in/raghava/gpcrpred/.

We have also suggested subfamilies for 42 sequences which were previously

identified as unclassified ClassA GPCRs. The supplementary information is

available at http://www.imtech.res.in/raghava/gpcrpred/info.html.

DOI: 10.1093/nar/gkh416

PMCID: PMC441554

PMID: 15215416 [Indexed for MEDLINE]

3319. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W380-2.

PRED-GPCR: GPCR recognition and family classification server.

Papasaikas PK(1), Bagos PG, Litou ZI, Promponas VJ, Hamodrakas SJ.

Author information:

(1)Department of Cell Biology and Biophysics, Faculty of Biology, University of

Athens, Panepistimiopolis, Athens 157 01, Greece.

The vast cell-surface receptor family of G-protein coupled receptors (GPCRs) is

the focus of both academic and pharmaceutical research due to their key role in

cell physiology along with their amenability to drug intervention. As the data

flow rate from the various genome and proteome projects continues to grow, so

does the need for fast, automated and reliable screening for new members of the

various GPCR families. PRED-GPCR is a free Internet service for GPCR recognition

and classification at the family level. A submitted sequence or set of sequences,

is queried against the PRED-GPCR library, housing 265 signature profile HMMs

corresponding to 67 well-characterized GPCR families. Users query the server

through a web interface and results are presented in HTML output format. The

server returns all single-motif matches along with the combined results for the

corresponding families. The service is available online since October 2003 at

http://bioinformatics.biol.uoa.gr/PRED-GPCR.

DOI: 10.1093/nar/gkh431

PMCID: PMC441569

PMID: 15215415 [Indexed for MEDLINE]

3320. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W372-4.

MITOPRED: a web server for the prediction of mitochondrial proteins.

Guda C(1), Guda P, Fahy E, Subramaniam S.

Author information:

(1)San Diego Supercomputer Center, University of California, San Diego, 9500

Gilman Drive, La Jolla, CA 92093, USA.

MITOPRED web server enables prediction of nucleus-encoded mitochondrial proteins

in all eukaryotic species. Predictions are made using a new algorithm based

primarily on Pfam domain occurrence patterns in mitochondrial and

non-mitochondrial locations. Pre-calculated predictions are instantly accessible

for proteomes of Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila,

Homo sapiens, Mus musculus and Arabidopsis species as well as all the eukaryotic

sequences in the Swiss-Prot and TrEMBL databases. Queries, at different

confidence levels, can be made through four distinct options: (i) entering

Swiss-Prot/TrEMBL accession numbers; (ii) uploading a local file with such

accession numbers; (iii) entering protein sequences; (iv) uploading a local file

containing protein sequences in FASTA format. Automated updates are scheduled for

the pre-calculated prediction database so as to provide access to the most

current data. The server, its documentation and the data are available from

http://mitopred.sdsc.edu.

DOI: 10.1093/nar/gkh374

PMCID: PMC441512

PMID: 15215413 [Indexed for MEDLINE]

3321. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W360-4.

GDAP: a web tool for genome-wide protein disulfide bond prediction.

O'Connor BD(1), Yeates TO.

Author information:

(1)UCLA-DOE Institute for Genomics and Proteomics, UCLA Molecular Biology

Institute and UCLA Department of Chemistry and Biochemistry, Los Angeles, USA.

The Genomic Disulfide Analysis Program (GDAP) provides web access to

computationally predicted protein disulfide bonds for over one hundred microbial

genomes, including both bacterial and achaeal species. In the GDAP process,

sequences of unknown structure are mapped, when possible, to known homologous

Protein Data Bank (PDB) structures, after which specific distance criteria are

applied to predict disulfide bonds. GDAP also accepts user-supplied protein

sequences and subsequently queries the PDB sequence database for the best

matches, scans for possible disulfide bonds and returns the results to the

client. These predictions are useful for a variety of applications and have

previously been used to show a dramatic preference in certain thermophilic

archaea and bacteria for disulfide bonds within intracellular proteins. Given the

central role these stabilizing, covalent bonds play in such organisms, the

predictions available from GDAP provide a rich data source for designing

site-directed mutants with more stable thermal profiles. The GDAP web application

is a gateway to this information and can be used to understand the role disulfide

bonds play in protein stability both in these unusual organisms and in sequences

of interest to the individual researcher. The prediction server can be accessed

at http://www.doe-mbi.ucla.edu/Services/GDAP.

DOI: 10.1093/nar/gkh376

PMCID: PMC441514

PMID: 15215411 [Indexed for MEDLINE]

3322. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W356-9.

SDPMOD: an automated comparative modeling server for small disulfide-bonded

proteins.

Kong L(1), Lee BT, Tong JC, Tan TW, Ranganathan S.

Author information:

(1)Department of Biochemistry, National University of Singapore, 8 Medical Drive,

117597, Singapore.

Small disulfide-bonded proteins (SDPs) are rich sources for therapeutic drugs.

Designing drugs from these proteins requires three-dimensional structural

information, which is only available for a subset of these proteins. SDPMOD

addresses this deficit in structural information by providing a freely available

automated comparative modeling service to the research community. For expert

users, SDPMOD offers a manual mode that permits the selection of a desired

template as well as a semi-automated mode that allows users to select the

template from a suggested list. Besides the selection of templates, expert users

can edit the target-template alignment, thus allowing further customization of

the modeling process. Furthermore, the web service provides model stereochemical

quality evaluation using PROCHECK. SDPMOD is freely accessible to academic users

via the web interface at http://proline.bic.nus.edu.sg/sdpmod.

DOI: 10.1093/nar/gkh394

PMCID: PMC441532

PMID: 15215410 [Indexed for MEDLINE]

3323. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W336-9.

Phydbac2: improved inference of gene function using interactive phylogenomic

profiling and chromosomal location analysis.

Enault F(1), Suhre K, Poirot O, Abergel C, Claverie JM.

Author information:

(1)Information Génomique & Structurale (UPR CNRS 2589), Institut de Biologie

Structurale et Microbiologie, 31, chemin Joseph Aiguier, 13402 Marseille Cedex

20, France. enault@igs.cnrs-mrs.fr

Phydbac (phylogenomic display of bacterial genes) implemented a method of

phylogenomic profiling using a distance measure based on normalized BLAST scores.

This method was able to increase the predictive power of phylogenomic profiling

by about 25% when compared to the classical approach based on Hamming distances.

Here we present a major extension of Phydbac (named here Phydbac2), that extends

both the concept and the functionality of the original web-service. While

phylogenomic profiles remain the central focus of Phydbac2, it now integrates

chromosomal proximity and gene fusion analyses as two additional

non-similarity-based indicators for inferring pairwise gene functional

relationships. Moreover, all presently available (January 2004) fully sequenced

bacterial genomes and those of three lower eukaryotes are now included in the

profiling process, thus increasing the initial number of reference genomes (71 in

Phydbac) to 150 in Phydbac2. Using the KEGG metabolic pathway database as a

benchmark, we show that the predictive power of Phydbac2 is improved by 27% over

the previous version. This gain is accounted for on one hand, by the increased

number of reference genomes (11%) and on the other hand, as a result of including

chromosomal proximity into the distance measure (16%). The expanded functionality

of Phydbac2 now allows the user to query more than 50 different genomes,

including at least one member of each major bacterial group, most major pathogens

and potential bio-terrorism agents. The search for co-evolving genes based on

consensus profiles from multiple organisms, the display of Phydbac2 profiles side

by side with COG information, the inclusion of KEGG metabolic pathway maps the

production of chromosomal proximity maps, and the possibility of collecting and

processing results from different Phydbac queries in a common shopping cart are

the main new features of Phydbac2. The Phydbac2 web server is available at

http://igs-server.cnrs-mrs.fr/phydbac/.

DOI: 10.1093/nar/gkh365

PMCID: PMC441503

PMID: 15215406 [Indexed for MEDLINE]

3324. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W332-5.

MyHits: a new interactive resource for protein annotation and domain

identification.

Pagni M(1), Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Falquet L.

Author information:

(1)Swiss Institute of Bioinformatics, CH-1066 Epalinges/Lausanne, Switzerland.

Marco.Pagni@isb-sib.ch

The MyHits web server (http://myhits.isb-sib.ch) is a new integrated service

dedicated to the annotation of protein sequences and to the analysis of their

domains and signatures. Guest users can use the system anonymously, with full

access to (i) standard bioinformatics programs (e.g. PSI-BLAST, ClustalW,

T-Coffee, Jalview); (ii) a large number of protein sequence databases, including

standard (Swiss-Prot, TrEMBL) and locally developed databases (splice variants);

(iii) databases of protein motifs (Prosite, Interpro); (iv) a precomputed list of

matches ('hits') between the sequence and motif databases. All databases are

updated on a weekly basis and the hit list is kept up to date incrementally. The

MyHits server also includes a new collection of tools to generate graphical

representations of pairwise and multiple sequence alignments including their

annotated features. Free registration enables users to upload their own sequences

and motifs to private databases. These are then made available through the same

web interface and the same set of analytical tools. Registered users can manage

their own sequences and annotations using only web tools and freeze their data in

their private database for publication purposes.

DOI: 10.1093/nar/gkh479

PMCID: PMC441617

PMID: 15215405 [Indexed for MEDLINE]

3325. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W321-6.

The PredictProtein server.

Rost B(1), Yachdav G, Liu J.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 650 West 168th Street BB217, New York, NY 10032, USA.

rost@columbia.edu

PredictProtein (http://www.predictprotein.org) is an Internet service for

sequence analysis and the prediction of protein structure and function. Users

submit protein sequences or alignments; PredictProtein returns multiple sequence

alignments, PROSITE sequence motifs, low-complexity regions (SEG), nuclear

localization signals, regions lacking regular structure (NORS) and predictions of

secondary structure, solvent accessibility, globular regions, transmembrane

helices, coiled-coil regions, structural switch regions, disulfide-bonds,

sub-cellular localization and functional annotations. Upon request fold

recognition by prediction-based threading, CHOP domain assignments, predictions

of transmembrane strands and inter-residue contacts are also available. For all

services, users can submit their query either by electronic mail or interactively

via the World Wide Web.

DOI: 10.1093/nar/gkh377

PMCID: PMC441515

PMID: 15215403 [Indexed for MEDLINE]

3326. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W318-20.

PredictRegulon: a web server for the prediction of the regulatory protein binding

sites and operons in prokaryote genomes.

Yellaboina S(1), Seshadri J, Kumar MS, Ranjan A.

Author information:

(1)Computational & Functional Genomics Group, Centre for DNA Fingerprinting and

Diagnostics, EMBnet India Node, Hyderabad 500076, India.

An interactive web server is developed for predicting the potential binding sites

and its target operons for a given regulatory protein in prokaryotic genomes. The

program allows users to submit known or experimentally determined binding sites

of a regulatory protein as ungapped multiple sequence alignments. It analyses the

upstream regions of all genes in a user-selected prokaryote genome and returns

the potential binding sites along with the downstream co-regulated genes

(operons). The known binding sites of a regulatory protein can also be used to

identify its orthologue binding sites in phylogeneticaly related genomes where

the trans-acting regulator protein and cognate cis-acting DNA sequences could be

conserved. PredictRegulon can be freely accessed from a link on our world wide

web server: http://www.cdfd.org.in/predictregulon/.

DOI: 10.1093/nar/gkh364

PMCID: PMC441502

PMID: 15215402 [Indexed for MEDLINE]

3327. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W313-7.

GOblet: a platform for Gene Ontology annotation of anonymous sequence data.

Groth D(1), Lehrach H, Hennig S.

Author information:

(1)Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin,

Germany.

GOblet is a comprehensive web server application providing the annotation of

anonymous sequence data with Gene Ontology (GO) terms. It uses a variety of

different protein databases (human, murines, invertebrates, plants, sp-trembl)

and their respective GO mappings. The user selects the appropriate database and

alignment threshold and thereafter submits single or multiple nucleotide or

protein sequences. Results are shown in different ways, e.g. as survey statistics

for the main GO categories for all sequences or as detailed results for each

single sequence that has been submitted. In its newest version, GOblet allows the

batch submission of sequences and provides an improved display of results with

the aid of Java applets. All output data, together with the Java applet, are

packed to a downloadable archive for local installation and analysis. GOblet can

be accessed freely at http://goblet.molgen.mpg.de.

DOI: 10.1093/nar/gkh406

PMCID: PMC441544

PMID: 15215401 [Indexed for MEDLINE]

3328. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W309-12.

AUGUSTUS: a web server for gene finding in eukaryotes.

Stanke M(1), Steinkamp R, Waack S, Morgenstern B.

Author information:

(1)University of Göttingen, Institut für Mikrobiologie und Genetik,

Goldschmidtstrasse 1, 37077 Göttingen, Germany. mstanke@gwdg.de

We present a www server for AUGUSTUS, a novel software program for ab initio gene

prediction in eukaryotic genomic sequences. Our method is based on a generalized

Hidden Markov Model with a new method for modeling the intron length

distribution. This method allows approximation of the true intron length

distribution more accurately than do existing programs. For genomic sequence data

from human and Drosophila melanogaster, the accuracy of AUGUSTUS is superior to

existing gene-finding approaches. The advantage of our program becomes apparent

especially for larger input sequences containing more than one gene. The server

is available at http://augustus.gobics.de.

DOI: 10.1093/nar/gkh379

PMCID: PMC441517

PMID: 15215400 [Indexed for MEDLINE]

3329. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W305-8.

AGenDA: gene prediction by cross-species sequence comparison.

Taher L(1), Rinner O, Garg S, Sczyrba A, Morgenstern B.

Author information:

(1)International Graduate School for Bioinformatics and Genome Research,

University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany.

Automatic gene prediction is one of the major challenges in computational

sequence analysis. Traditional approaches to gene finding rely on statistical

models derived from previously known genes. By contrast, a new class of

comparative methods relies on comparing genomic sequences from evolutionary

related organisms to each other. These methods are based on the concept of

phylogenetic footprinting: they exploit the fact that functionally important

regions in genomic sequences are usually more conserved than non-functional

regions. We created a WWW-based software program for homology-based gene

prediction at BiBiServ (Bielefeld Bioinformatics Server). Our tool takes pairs of

evolutionary related genomic sequences as input data, e.g. from human and mouse.

The server runs CHAOS and DIALIGN to create an alignment of the input sequences

and subsequently searches for conserved splicing signals and start/stop codons

near regions of local sequence conservation. Genes are predicted based on local

homology information and splice signals. The server returns predicted genes

together with a graphical representation of the underlying alignment. The program

is available at http://bibiserv.TechFak.Uni-Bielefeld.DE/agenda/.

DOI: 10.1093/nar/gkh386

PMCID: PMC441524

PMID: 15215399 [Indexed for MEDLINE]

3330. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W301-4.

e2g: an interactive web-based server for efficiently mapping large EST and cDNA

sets to genomic sequences.

Krüger J(1), Sczyrba A, Kurtz S, Giegerich R.

Author information:

(1)Technische Fakultät, Universität Bielefeld, D-33594 Bielefeld, Germany.

e2g is a web-based server which efficiently maps large expressed sequence tag

(EST) and cDNA datasets to genomic DNA. It significantly extends the volume of

data that can be mapped in reasonable time, and makes this improved efficiency

available as a web service. Our server hosts large collections of EST sequences

(e.g. 4.1 million mouse ESTs of 1.87 Gb) in precomputed indexed data structures

for efficient sequence comparison. The user can upload a genomic DNA sequence of

interest and rapidly compare this to the complete collection of ESTs on the

server. This delivers a mapping of the ESTs on the genomic DNA. The e2g web

interface provides a graphical overview of the mapping. Alignments of the mapped

EST regions with parts of the genomic sequence are visualized. Zooming functions

allow the user to interactively explore the results. Mapped sequences can be

downloaded for further analysis. e2g is available on the Bielefeld University

Bioinformatics Server at http://bibiserv.techfak.uni-bielefeld.de/e2g/.

DOI: 10.1093/nar/gkh478

PMCID: PMC441616

PMID: 15215398 [Indexed for MEDLINE]

3331. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W293-300.

GFINDer: Genome Function INtegrated Discoverer through dynamic annotation,

statistical analysis, and mining.

Masseroli M(1), Martucci D, Pinciroli F.

Author information:

(1)Bioengineering Department, Politecnico di Milano, I-20133 Milano, Italy.

masseroli@biomed.polimi.it

Statistical and clustering analyses of gene expression results from high-density

microarray experiments produce lists of hundreds of genes regulated

differentially, or with particular expression profiles, in the conditions under

study. Independent of the microarray platforms and analysis methods used, these

lists must be biologically interpreted to gain a better knowledge of the

patho-physiological phenomena involved. To this end, numerous biological

annotations are available within heterogeneous and widely distributed databases.

Although several tools have been developed for annotating lists of genes, most of

them do not give methods for evaluating the relevance of the annotations

provided, or for estimating the functional bias introduced by the gene set on the

array used to identify the gene list considered. We developed Genome Functional

INtegrated Discoverer (GFINDer), a web server able to automatically provide

large-scale lists of user-classified genes with functional profiles biologically

characterizing the different gene classes in the list. GFINDer automatically

retrieves annotations of several functional categories from different sources,

identifies the categories enriched in each class of a user-classified gene list

and calculates statistical significance values for each category. Moreover,

GFINDer enables the functional classification of genes according to mined

functional categories and the statistical analysis is of the classifications

obtained, aiding better interpretation of microarray experiment results. GFINDer

is available online at http://www.medinfopoli.polimi.it/GFINDer/.

DOI: 10.1093/nar/gkh432

PMCID: PMC441570

PMID: 15215397 [Indexed for MEDLINE]

3332. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W273-9.

VISTA: computational tools for comparative genomics.

Frazer KA(1), Pachter L, Poliakov A, Rubin EM, Dubchak I.

Author information:

(1)Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, CA 94043, USA.

Comparison of DNA sequences from different species is a fundamental method for

identifying functional elements in genomes. Here, we describe the VISTA family of

tools created to assist biologists in carrying out this task. Our first VISTA

server at http://www-gsd.lbl.gov/vista/ was launched in the summer of 2000 and

was designed to align long genomic sequences and visualize these alignments with

associated functional annotations. Currently the VISTA site includes multiple

comparative genomics tools and provides users with rich capabilities to browse

pre-computed whole-genome alignments of large vertebrate genomes and other groups

of organisms with VISTA Browser, to submit their own sequences of interest to

several VISTA servers for various types of comparative analysis and to obtain

detailed comparative analysis results for a set of cardiovascular genes. We

illustrate capabilities of the VISTA site by the analysis of a 180 kb interval on

human chromosome 5 that encodes for the kinesin family member 3A (KIF3A) protein.

DOI: 10.1093/nar/gkh458

PMCID: PMC441596

PMID: 15215394 [Indexed for MEDLINE]

3333. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W253-6.

CREME: Cis-Regulatory Module Explorer for the human genome.

Sharan R(1), Ben-Hur A, Loots GG, Ovcharenko I.

Author information:

(1)International Computer Science Institute, 1947 Center Street, Berkeley, CA

94704, USA. roded@icsi.berkeley.edu

The binding of transcription factors to specific regulatory sequence elements is

a primary mechanism for controlling gene transcription. Eukaryotic genes are

often regulated by several transcription factors whose binding sites are tightly

clustered and form cis-regulatory modules. In this paper, we present a web

server, CREME, for identifying and visualizing cis-regulatory modules in the

promoter regions of a given set of potentially co-regulated genes. CREME relies

on a database of putative transcription factor binding sites that have been

annotated across the human genome using a library of position weight matrices and

evolutionary conservation with the mouse and rat genomes. A search algorithm is

applied to this data set to identify combinations of transcription factors whose

binding sites tend to co-occur in close proximity in the promoter regions of the

input gene set. The identified cis-regulatory modules are statistically scored

and significant combinations are reported and graphically visualized. Our web

server is available at http://creme.dcode.org.

DOI: 10.1093/nar/gkh385

PMCID: PMC441523

PMID: 15215390 [Indexed for MEDLINE]

3334. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W217-21.

rVISTA 2.0: evolutionary analysis of transcription factor binding sites.

Loots GG(1), Ovcharenko I.

Author information:

(1)Genome Biology Division and EEBI Computing Division, Lawrence Livermore

National Laboratory, 7000 East Avenue, L-441, Livermore, CA 94550, USA.

Identifying and characterizing the transcription factor binding site (TFBS)

patterns of cis-regulatory elements represents a challenge, but holds promise to

reveal the regulatory language the genome uses to dictate transcriptional

dynamics. Several studies have demonstrated that regulatory modules are under

positive selection and, therefore, are often conserved between related species.

Using this evolutionary principle, we have created a comparative tool, rVISTA,

for analyzing the regulatory potential of noncoding sequences. Our ability to

experimentally identify functional noncoding sequences is extremely limited,

therefore, rVISTA attempts to fill this great gap in genomic analysis by offering

a powerful approach for eliminating TFBSs least likely to be biologically

relevant. The rVISTA tool combines TFBS predictions, sequence comparisons and

cluster analysis to identify noncoding DNA regions that are evolutionarily

conserved and present in a specific configuration within genomic sequences. Here,

we present the newly developed version 2.0 of the rVISTA tool, which can process

alignments generated by both the zPicture and blastz alignment programs or use

pre-computed pairwise alignments of several vertebrate genomes available from the

ECR Browser and GALA database. The rVISTA web server is closely interconnected

with the TRANSFAC database, allowing users to either search for matrices present

in the TRANSFAC library collection or search for user-defined consensus

sequences. The rVISTA tool is publicly available at http://rvista.dcode.org/.

DOI: 10.1093/nar/gkh383

PMCID: PMC441521

PMID: 15215384 [Indexed for MEDLINE]

3335. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W195-8.

MSCAN: identification of functional clusters of transcription factor binding

sites.

Alkema WB(1), Johansson O, Lagergren J, Wasserman WW.

Author information:

(1)Center for Genomics and Bioinformatics, Karolinska Institutet, SE-17177

Stockholm, Sweden.

Identification of functional transcription factor binding sites in genomic

sequences is notoriously difficult. The critical problem is the low specificity

of predictions, which directly reflects the low target specificity of DNA binding

proteins. To overcome the noise produced in predictions of individual binding

sites, a new generation of algorithms achieves better predictive specificity by

focusing on locally dense clusters of binding sites. MSCAN is a leading method

for binding site cluster detection that determines the significance of observed

sites while correcting for local compositional bias of sequences. The algorithm

is highly flexible, applying any set of input binding models to the analysis of a

user-specified sequence. From the user's perspective, a key feature of the system

is that no reference data sets of regulatory sequences from co-regulated genes

are required to train the algorithm. The output from MSCAN consists of an ordered

list of sequence segments that contain potential regulatory modules. We have

chosen the features in MSCAN such that sequence and matrix retrieval is highly

facilitated, resulting in a web server that is intuitive to use. MSCAN is

available at http://mscan.cgb.ki.se/cgi-bin/MSCAN.

DOI: 10.1093/nar/gkh387

PMCID: PMC441525

PMID: 15215379 [Indexed for MEDLINE]

3336. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W173-5.

The sequencing-based typing tool of dbMHC: typing highly polymorphic gene

sequences.

Helmberg W(1), Dunivin R, Feolo M.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, Building 45, 8600 Rockville Pike, Bethesda, MD

20894, USA. helmberg@ncbi.nlm.nih.gov

The dbMHC resource (http://www.ncbi.nlm.nih.gov/mhc/sbt.cgi?cmd=main) at the

National Center for Biotechnology Information (NCBI) has developed an online tool

for evaluating the allelic composition of sequencing-based typing (SBT) results

of cDNA or genomic sequences. Whether the samples are heterozygous, haploid or a

combination of the two, they can be compared with two up-to-date databases of all

known alleles of several human leukocyte antigen (HLA) and killer cell

immunoglobulin-like receptor (KIR) loci. The results of the submission are

returned as a table of potential allele hits, along with the respective base

changes and an interactive sequence viewer for close examination of the

alignment.

DOI: 10.1093/nar/gkh424

PMCID: PMC441562

PMID: 15215374 [Indexed for MEDLINE]

3337. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W170-2.

SNPbox: web-based high-throughput primer design from gene to genome.

Weckx S(1), De Rijk P, Van Broeckhoven C, Del-Favero J.

Author information:

(1)Department of Molecular Genetics, Bioinformatics Unit, Flanders

Interuniversity Institute for Biotechnology, University of Antwerp, Belgium.

SNPbox is a modular software package that automates the design of PCR primers for

large-scale amplification and sequencing projects in a standardized manner

resulting in high-quality PCR amplicons with a low failure rate. Here, we present

the SNPbox web server at http://www.SNPbox.org, which hosts the SNPbox web

service as well as the data from SNPbox analysis of all Ensembl exons. The data

of this genome-wide SNPbox application can be visualized in Ensembl's ContigView

through a DAS (distributed annotation system) annotation server.

DOI: 10.1093/nar/gkh369

PMCID: PMC441507

PMID: 15215373 [Indexed for MEDLINE]

3338. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W166-9.

PDA: a pipeline to explore and estimate polymorphism in large DNA databases.

Casillas S(1), Barbadilla A.

Author information:

(1)Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona,

08193 Bellaterra, Barcelona, Spain.

Polymorphism studies are one of the main research areas of this genomic era. To

date, however, no available web server or software package has been designed to

automate the process of exploring and estimating nucleotide polymorphism in large

DNA databases. Here, we introduce a novel software, PDA, Pipeline Diversity

Analysis, that automatically can (i) search for polymorphic sequences in large

databases, and (ii) estimate their genetic diversity. PDA is a collection of

modules, mainly written in Perl, which works sequentially as follows: unaligned

sequence retrieved from a DNA database are automatically classified by organism

and gene, and aligned using the ClustalW algorithm. Sequence sets are regrouped

depending on their similarity scores. Main diversity parameters, including

polymorphism, synonymous and non-synonymous substitutions, linkage disequilibrium

and codon bias are estimated both for the full length of the sequences and for

specific functional regions. Program output includes a database with all

sequences and estimations, and HTML pages with summary statistics, the performed

alignments and a histogram maker tool. PDA is an essential tool to explore

polymorphism in large DNA databases for sequences from different genes,

populations or species. It has already been successfully applied to create a

secondary database. PDA is available on the web at http://pda.uab.es/.

DOI: 10.1093/nar/gkh428

PMCID: PMC441566

PMID: 15215372 [Indexed for MEDLINE]

3339. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W160-5.

The ERPIN server: an interface to profile-based RNA motif identification.

Lambert A(1), Fontaine JF, Legendre M, Leclerc F, Permal E, Major F, Putzer H,

Delfour O, Michot B, Gautheret D.

Author information:

(1)CNRS UMR 6207, Université de la Méditerranée, Luminy Case 906, 13288

Marseille, Cedex 09, France.

ERPIN is an RNA motif identification program that takes an RNA sequence alignment

as an input and identifies related sequences using a profile-based dynamic

programming algorithm. ERPIN differs from other RNA motif search programs in its

ability to capture subtle biases in the training set and produce highly specific

and sensitive searches, while keeping CPU requirements at a practical level. In

its latest version, ERPIN also computes E-values, which tell biologists how

likely they are to encounter a specific sequence match by chance-a useful

indication of biological significance. We present here the ERPIN online search

interface (http://tagc.univ-mrs.fr/erpin/). This web server automatically

performs ERPIN searches for different RNA genes or motifs, using predefined

training sets and search parameters. With a couple of clicks, users can analyze

an entire bacterial genome or a genomic segment of up to 5Mb for the presence of

tRNAs, 5S rRNAs, SRP RNA, C/D box snoRNAs, hammerhead motifs, miRNAs and other

motifs. Search results are displayed with sequence, score, position, E-value and

secondary structure graphics. An example of a complete genome scan is provided,

as well as an evaluation of run times and specificity/sensitivity information for

all available motifs.

DOI: 10.1093/nar/gkh418

PMCID: PMC441556

PMID: 15215371 [Indexed for MEDLINE]

3340. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W154-9.

Riboswitch finder--a tool for identification of riboswitch RNAs.

Bengert P(1), Dandekar T.

Author information:

(1)Biochemiezentrum, Project Group B2 (Dandekar/Schirmer), University of

Heidelberg, Im Neuenheimer Feld 504, 69120 Heidelberg, Germany.

We describe a dedicated RNA motif search program and web server to identify RNA

riboswitches. The Riboswitch finder analyses a given sequence using the web

interface, checks specific sequence elements and secondary structure, calculates

and displays the energy folding of the RNA structure and runs a number of tests

including this information to determine whether high-sensitivity riboswitch

motifs (or variants) according to the Bacillus subtilis type are present in the

given RNA sequence. Batch-mode determination (all sequences input at once and

separated by FASTA format) is also possible. The program has been implemented and

is available both as local software for in-house installation and as a web server

at http://www.biozentrum.uni-wuerzburg.de/bioinformatik/Riboswitch/.

DOI: 10.1093/nar/gkh352

PMCID: PMC441490

PMID: 15215370 [Indexed for MEDLINE]

3341. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W150-3.

RDfolder: a web server for prediction of RNA secondary structure.

Ying X(1), Luo H, Luo J, Li W.

Author information:

(1)Beijing Institute of Basic Medical Sciences, PO Box 130(3), Beijing 100850,

People's Republic of China.

Prediction of RNA secondary structure is important in the functional analysis of

RNA molecules. The RDfolder web server described in this paper provides two

methods for prediction of RNA secondary structure: random stacking of helical

regions and helical regions distribution. The random stacking method predicts

secondary structure by Monte Carlo simulations. The method of helical regions

distribution predicts secondary structure based on the helices that appear most

frequently in the set of structures, which are generated by the random stacking

method. The RDfolder web server can be accessed at http://rna.cbi.pku.edu.cn.

DOI: 10.1093/nar/gkh445

PMCID: PMC441583

PMID: 15215369 [Indexed for MEDLINE]

3342. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W146-9.

ILM: a web server for predicting RNA secondary structures with pseudoknots.

Ruan J(1), Stormo GD, Zhang W.

Author information:

(1)Department of Computer Science and Engineering, Washington University in St

Louis, St Louis, MO 63130, USA.

The ILM web server provides a web interface to two algorithms, iterated loop

matching and maximum weighted matching, for efficiently predicting RNA secondary

structures with pseudoknots. The algorithms can utilize either thermodynamic or

comparative information or both, and thus can work on both aligned and individual

sequences. Predicted secondary structures are presented in several formats

compatible with a variety of existing visualization tools. The service can be

accessed at http://cic.cs.wustl.edu/RNA/.

DOI: 10.1093/nar/gkh444

PMCID: PMC441582

PMID: 15215368 [Indexed for MEDLINE]

3343. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W142-5.

CARNAC: folding families of related RNAs.

Touzet H(1), Perriquet O.

Author information:

(1)Laboratoire d'Informatique Fondamentale de Lille, UMR CNRS 8022, Université

des Sciences et Technologies de Lille, France.

We present a tool for the prediction of conserved secondary structure elements of

a family of homologous non-coding RNAs. Our method does not require any prior

multiple sequence alignment. Thus, it successfully applies to datasets with low

primary structure similarity. The functionality is demonstrated using three

example datasets: sequences of RNase P RNAs, ciliate telomerases and enterovirus

messenger RNAs. CARNAC has a web server that can be accessed at the URL

http://bioinfo.lifl.fr/carnac.

DOI: 10.1093/nar/gkh415

PMCID: PMC441553

PMID: 15215367 [Indexed for MEDLINE]

3344. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W135-41.

Sfold web server for statistical folding and rational design of nucleic acids.

Ding Y(1), Chan CY, Lawrence CE.

Author information:

(1)Bioinformatics Center, Wadsworth Center, New York State Department of Health,

150 New Scotland Avenue, Albany, NY 12208, USA. yding@wadsworth.org

The Sfold web server provides user-friendly access to Sfold, a recently developed

nucleic acid folding software package, via the World Wide Web (WWW). The software

is based on a new statistical sampling paradigm for the prediction of RNA

secondary structure. One of the main objectives of this software is to offer

computational tools for the rational design of RNA-targeting nucleic acids, which

include small interfering RNAs (siRNAs), antisense oligonucleotides and

trans-cleaving ribozymes for gene knock-down studies. The methodology for siRNA

design is based on a combination of RNA target accessibility prediction, siRNA

duplex thermodynamic properties and empirical design rules. Our approach to

target accessibility evaluation is an original extension of the underlying RNA

folding algorithm to account for the likely existence of a population of

structures for the target mRNA. In addition to the application modules Sirna,

Soligo and Sribo for siRNAs, antisense oligos and ribozymes, respectively, the

module Srna offers comprehensive features for statistical representation of

sampled structures. Detailed output in both graphical and text formats is

available for all modules. The Sfold server is available at

http://sfold.wadsworth.org and http://www.bioinfo.rpi.edu/applications/sfold.

DOI: 10.1093/nar/gkh449

PMCID: PMC441587

PMID: 15215366 [Indexed for MEDLINE]

3345. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W130-4.

siRNA Selection Server: an automated siRNA oligonucleotide prediction server.

Yuan B(1), Latek R, Hossbach M, Tuschl T, Lewitter F.

Author information:

(1)Whitehead Institute for Biomedical Research, Bioinformatics and Research

Computing, Nine Cambridge Center, Cambridge, MA 02142, USA.

The Whitehead siRNA (short interfering RNA) Selection Web Server

(http://jura.wi.mit.edu/bioc/siRNA) automates the design of short

oligonucleotides that can specifically 'knock down' expression of target genes.

These short sequences are about 21 nt in length, and when synthesized as double

stranded RNA and introduced into cell culture, can reduce or eliminate the

function of the target gene. Depending on the length of a gene, there are

potentially numerous combinations of possible 21mers. Some experimental evidence

has already shown that not all 21mers in a gene have the same effectiveness at

silencing gene function. Our tool incorporates published design rules and

presents the scientist with information about uniqueness of the 21mers within the

genome, thermodynamic stability of the double stranded RNA duplex, GC content,

presence of SNPs and other features that may contribute to the effectiveness of a

siRNA.

DOI: 10.1093/nar/gkh366

PMCID: PMC441504

PMID: 15215365 [Indexed for MEDLINE]

3346. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W108-12.

A web server for performing electronic PCR.

Rotmistrovsky K(1), Jang W, Schuler GD.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, Bethesda, MD 20984, USA.

'Electronic PCR' (e-PCR) refers to a computational procedure that is used to

search DNA sequences for sequence tagged sites (STSs), each of which is defined

by a pair of primer sequences and an expected PCR product size. To gain speed,

our implementation extracts short 'words' from the 3' end of each primer and

stores them in a sorted hash table that can be accessed efficiently during the

search. One recent improvement is the use of overlapping discontinuous words to

allow matches to be found despite the presence of a mismatch. Moreover, it is

possible to allow gaps in the alignment between the primer and the sequence. The

effect of these changes is to improve sensitivity without significantly affecting

specificity. The new software provides a search mode using a query STS against a

sequence database to augment the previously available mode using a query sequence

against an STS database. Finally, e-PCR may now be used through a web service,

with search results linked to other web resources such as the UniSTS database and

the MapViewer genome browser. The e-PCR web server may be found at

www.ncbi.nlm.nih.gov/sutils/e-pcr.

DOI: 10.1093/nar/gkh450

PMCID: PMC441588

PMID: 15215361 [Indexed for MEDLINE]

3347. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W104-7.

Qgrid: clustering tool for detecting charged and hydrophobic regions in proteins.

Ahmad S(1), Sarai A.

Author information:

(1)Department of Biochemical Science and Engineering, Kyushu Institute of

Technology, Iizuka 820 8502, Fukuoka-ken, Japan. shandar@bse.kyutech.ac.jp

We have developed a simple but powerful method and web server to quickly locate

charged and hydrophobic clusters in proteins

(http://www.netasa.org/qgrid/index.html). For the charged clusters, each atom in

the protein is first assigned a charge according to a standard force field. Then

a box is created with dimensions corresponding to the range of atomic

coordinates. This box is then divided into cubic grids of selected size, which

now have one or more charged atoms in them. This leaves each grid with a certain

amount of charge. Cubic grids with more than a cutoff charge are then clustered

using a hierarchical clustering method based on Euclidean distance. A tree

diagram made from the resulting clusters indicates the distribution of charged

and hydrophobic regions of the protein. Hydrophobic clusters are developed by

grouping the positions of C(alpha) atoms of such residues. We propose that such a

tree representation will be helpful in detecting protein-protein interfaces,

structure similarity and motif detection.

DOI: 10.1093/nar/gkh363

PMCID: PMC441501

PMID: 15215360 [Indexed for MEDLINE]

3348. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W100-3.

CE-MC: a multiple protein structure alignment server.

Guda C(1), Lu S, Scheeff ED, Bourne PE, Shindyalov IN.

Author information:

(1)San Diego Supercomputer Center, University of California, San Diego, 9500

Gilman Drive, La Jolla, CA 92093, USA. babu@sdsc.edu

CE-MC server (http://cemc.sdsc.edu) provides a web-based facility for the

alignment of multiple protein structures based on C-alpha coordinate distances,

using combinatorial extension (CE) and Monte Carlo (MC) optimization methods.

Alignments are possible for user-selected PDB (Protein Data Bank) chains as well

as for user-uploaded structures or the combination of the two. The whole process

of generating multiple structure alignments involves three distinct steps, i.e.

all-to-all pairwise alignment using the CE algorithm, iterative global

optimization of a multiple alignment using the MC algorithm and formatting MC

results using the JOY program. The server can be used to get multiple alignments

for up to 25 protein structural chains with the flexibility of uploading multiple

coordinate files and performing multiple structure alignment for user-selected

PDB chains. For large-scale jobs and local installation of the CE-MC program,

users can download the source code and precompiled binaries from the web server.

DOI: 10.1093/nar/gkh464

PMCID: PMC441602

PMID: 15215359 [Indexed for MEDLINE]

3349. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W96-9.

ClusPro: a fully automated algorithm for protein-protein docking.

Comeau SR(1), Gatchell DW, Vajda S, Camacho CJ.

Author information:

(1)Bioinformatics Graduate Program, Boston University, 44 Cummington Street,

Boston, MA 02215, USA.

ClusPro (http://nrc.bu.edu/cluster) represents the first fully automated,

web-based program for the computational docking of protein structures. Users may

upload the coordinate files of two protein structures through ClusPro's web

interface, or enter the PDB codes of the respective structures, which ClusPro

will then download from the PDB server (http://www.rcsb.org/pdb/). The docking

algorithms evaluate billions of putative complexes, retaining a preset number

with favorable surface complementarities. A filtering method is then applied to

this set of structures, selecting those with good electrostatic and desolvation

free energies for further clustering. The program output is a short list of

putative complexes ranked according to their clustering properties, which is

automatically sent back to the user via email.

DOI: 10.1093/nar/gkh354

PMCID: PMC441492

PMID: 15215358 [Indexed for MEDLINE]

3350. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W76-82.

ProteMiner-SSM: a web server for efficient analysis of similar protein tertiary

substructures.

Chang DT(1), Chen CY, Chung WC, Oyang YJ, Juan HF, Huang HC.

Author information:

(1)Department of Computer Science and Information Engineering, National Taiwan

University, Taipei, Taiwan, ROC.

Analysis of protein-ligand interactions is a fundamental issue in drug design. As

the detailed and accurate analysis of protein-ligand interactions involves

calculation of binding free energy based on thermodynamics and even quantum

mechanics, which is highly expensive in terms of computing time, conformational

and structural analysis of proteins and ligands has been widely employed as a

screening process in computer-aided drug design. In this paper, a web server

called ProteMiner-SSM designed for efficient analysis of similar protein tertiary

substructures is presented. In one experiment reported in this paper, the web

server has been exploited to obtain some clues about a biochemical hypothesis.

The main distinction in the software design of the web server is the filtering

process incorporated to expedite the analysis. The filtering process extracts the

residues located in the caves of the protein tertiary structure for analysis and

operates with O(nlogn) time complexity, where n is the number of residues in the

protein. In comparison, the alpha-hull algorithm, which is a widely used

algorithm in computer graphics for identifying those instances that are on the

contour of a three-dimensional object, features O(n2) time complexity.

Experimental results show that the filtering process presented in this paper is

able to speed up the analysis by a factor ranging from 3.15 to 9.37 times. The

ProteMiner-SSM web server can be found at http://proteminer.csie.ntu.edu.tw/.

There is a mirror site at http://p4.sbl.bc.sinica.edu.tw/proteminer/.

DOI: 10.1093/nar/gkh425

PMCID: PMC441563

PMID: 15215355 [Indexed for MEDLINE]

3351. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W73-5.

InterWeaver: interaction reports for discovering potential protein interaction

partners with online evidence.

Zhang Z(1), Ng SK.

Author information:

(1)Knowledge Discovery Department, Institute for Infocomm Research, 21 Heng Mui

Keng Terrace, Singapore 119613. zzhang@i2r.a-star.edu.sg

InterWeaver is a web server for discovering potential protein interactions with

online evidence automatically extracted from protein interaction databases,

literature abstracts, domain fusion events and domain interactions. Given a new

protein sequence, the server identifies potential interaction partners using two

approaches. In the homology-based approach, the system performs sequence homology

searches to find similar proteins in other species, and then searches the protein

interaction databases and the biomedical literature for interaction partners. In

the domain-based approach, the system detects the domains in the input protein

sequence and searches databases of domain fusion events and putative domain

interactions to suggest potential interacting partners. The results are compiled

into a personalized and downloadable interaction report to aid biologists in

their discovery of protein interactions. InterWeaver is freely available for

academic users at http://interweaver.i2r.a-star.edu.sg/.

DOI: 10.1093/nar/gkh437

PMCID: PMC441575

PMID: 15215354 [Indexed for MEDLINE]

3352. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W50-4.

Consensus alignment server for reliable comparative modeling with distant

templates.

Prasad JC(1), Vajda S, Camacho CJ.

Author information:

(1)Bioinformatics Graduate Program, Boston University, Boston, MA 02215, USA.

Consensus is a server developed to produce high-quality alignments for

comparative modeling, and to identify the alignment regions reliable for copying

from a given template. This is accomplished even when target-template sequence

identity is as low as 5%. Combining the output from five different alignment

methods, the server produces a consensus alignment, with a reliability measure

indicated for each position and a prediction of the regions suitable for

modeling. Models built using the server predictions are typically within 3 A rms

deviations from the crystal structure. Users can upload a target protein sequence

and specify a template (PDB code); if no template is given, the server will

search for one. The method has been validated on a large set of homologous

protein structure pairs. The Consensus server should prove useful for modelers

for whom the structural reliability of the model is critical in their

applications. It is currently available at

http://structure.bu.edu/cgi-bin/consensus/consensus.cgi.

DOI: 10.1093/nar/gkh456

PMCID: PMC441594

PMID: 15215349 [Indexed for MEDLINE]

3353. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W48-9.

WAViS server for handling, visualization and presentation of multiple alignments

of nucleotide or amino acids sequences.

Zika R(1), Paces J, Pavlícek A, Paces V.

Author information:

(1)Institute of Molecular Genetics, Academy of Sciences of the Czech Republic,

CZ-16637 Prague, Czech Republic.

Web Alignment Visualization Server contains a set of web-tools designed for quick

generation of publication-quality color figures of multiple alignments of

nucleotide or amino acids sequences. It can be used for identification of

conserved regions and gaps within many sequences using only common web browsers.

The server is accessible at http://wavis.img.cas.cz.

DOI: 10.1093/nar/gkh358

PMCID: PMC441496

PMID: 15215348 [Indexed for MEDLINE]

3354. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W41-4.

The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences.

Brudno M(1), Steinkamp R, Morgenstern B.

Author information:

(1)Department of Computer Science, Stanford University, Stanford, CA 94305, USA.

Cross-species sequence comparison is a powerful approach to analyze functional

sites in genomic sequences and many discoveries have been made based on genomic

alignments. Herein, we present a WWW-based software system for multiple alignment

of large genomic sequences. Our server utilizes the previously developed

combination of CHAOS and DIALIGN to achieve both speed and alignment accuracy.

CHAOS is a fast database search tool that creates a list of local sequence

similarities. These are used by DIALIGN as anchor points to speed up the final

alignment procedure. The resulting alignment is returned to the user in different

formats together with a list of anchor points found by CHAOS. The CHAOS/DIALIGN

software is freely available at

http://dialign.gobics.de/chaos-dialign-submission.

DOI: 10.1093/nar/gkh361

PMCID: PMC441499

PMID: 15215346 [Indexed for MEDLINE]

3355. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W37-40.

3DCoffee@igs: a web server for combining sequences and structures into a multiple

sequence alignment.

Poirot O(1), Suhre K, Abergel C, O'Toole E, Notredame C.

Author information:

(1)Information Génomique et Structurale UPR2589-CNRS, CNRS, 31, Chemin Joseph

Aiguier, 13 402 Marseille Cedex 20, France.

This paper presents 3DCoffee@igs, a web-based tool dedicated to the computation

of high-quality multiple sequence alignments (MSAs). 3D-Coffee makes it possible

to mix protein sequences and structures in order to increase the accuracy of the

alignments. Structures can be either provided as PDB identifiers or directly

uploaded into the server. Given a set of sequences and structures, pairs of

structures are aligned with SAP while sequence-structure pairs are aligned with

Fugue. The resulting collection of pairwise alignments is then combined into an

MSA with the T-Coffee algorithm. The server and its documentation are available

from http://igs-server.cnrs-mrs.fr/Tcoffee/.

DOI: 10.1093/nar/gkh382

PMCID: PMC441520

PMID: 15215345 [Indexed for MEDLINE]

3356. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W33-6.

DIALIGN: multiple DNA and protein sequence alignment at BiBiServ.

Morgenstern B(1).

Author information:

(1)University of Göttingen, Institute of Microbiology and Genetics,

Goldschmidtstr. 1, 37077 Göttingen, Germany. bmorgen@gwdg.de

DIALIGN is a widely used software tool for multiple DNA and protein sequence

alignment. The program combines local and global alignment features and can

therefore be applied to sequence data that cannot be correctly aligned by more

traditional approaches. DIALIGN is available online through Bielefeld

Bioinformatics Server (BiBiServ). The downloadable version of the program offers

several new program features. To compare the output of different alignment

programs, we developed the program AltAVisT. Our software is available at

http://bibiserv.TechFak.Uni-Bielefeld.DE/dialign/.

DOI: 10.1093/nar/gkh373

PMCID: PMC441511

PMID: 15215344 [Indexed for MEDLINE]

3357. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W16-9.

PubCrawler: keeping up comfortably with PubMed and GenBank.

Hokamp K(1), Wolfe KH.

Author information:

(1)Department of Molecular Biology and Biochemistry, Simon Fraser University,

Burnaby, BC, Canada.

The free PubCrawler web service (http://www.pubcrawler.ie) has been operating for

five years and so far has brought literature and sequence updates to over 22 000

users. It provides information on a personalized web page whenever new articles

appear in PubMed or when new sequences are found in GenBank that are specific to

customized queries. The server also acts as an automatic alerting system by

sending out short notifications or emails with the latest updates as soon as they

become available. A new output format and more flexibility for the email

formatting help PubCrawler cope with increasing challenges arising from browser

incompatibilities and mail filters, therefore making it suitable for a wide range

of users.

DOI: 10.1093/nar/gkh453

PMCID: PMC441591

PMID: 15215341 [Indexed for MEDLINE]

3358. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W10-5.

The web server of IBM's Bioinformatics and Pattern Discovery group: 2004 update.

Huynh T(1), Rigoutsos I.

Author information:

(1)Bioinformatics and Pattern Discovery group, IBM T.J. Watson Research Center,

PO Box 218, Yorktown Heights, NY 10598, USA.

In this report, we provide an update on the services and content which are

available on the web server of IBM's Bioinformatics and Pattern Discovery group.

The server, which is operational around the clock, provides access to a large

number of methods that have been developed and published by the group's members.

There is an increasing number of problems that these tools can help tackle; these

problems range from the discovery of patterns in streams of events and the

computation of multiple sequence alignments, to the discovery of genes in nucleic

acid sequences, the identification--directly from sequence--of structural

deviations from alpha-helicity and the annotation of amino acid sequences for

antimicrobial activity. Additionally, annotations for more than 130 archaeal,

bacterial, eukaryotic and viral genomes are now available on-line and can be

searched interactively. The tools and code bundles continue to be accessible from

http://cbcsrv.watson.ibm.com/Tspd.html whereas the genomics annotations are

available at http://cbcsrv.watson.ibm.com/Annotations/.

DOI: 10.1093/nar/gkh367

PMCID: PMC441505

PMID: 15215340 [Indexed for MEDLINE]

3359. Proteins. 2004 Jul 1;56(1):93-101.

A physical reference state unifies the structure-derived potential of mean force

for protein folding and binding.

Liu S(1), Zhang C, Zhou H, Zhou Y.

Author information:

(1)Howard Hughes Medical Institute Center for Single Molecule Biophysics,

Department of Physiology and Biophysics, State University of New York at Buffalo,

Buffalo, New York 14214, USA.

Extracting knowledge-based statistical potential from known structures of

proteins is proved to be a simple, effective method to obtain an approximate

free-energy function. However, the different compositions of amino acid residues

at the core, the surface, and the binding interface of proteins prohibited the

establishment of a unified statistical potential for folding and binding despite

the fact that the physical basis of the interaction (water-mediated interaction

between amino acids) is the same. Recently, a physical state of ideal gas, rather

than a statistically averaged state, has been used as the reference state for

extracting the net interaction energy between amino acid residues of monomeric

proteins. Here, we find that this monomer-based potential is more accurate than

an existing all-atom knowledge-based potential trained with interfacial

structures of dimers in distinguishing native complex structures from docking

decoys (100% success rate vs. 52% in 21 dimer/trimer decoy sets). It is also more

accurate than a recently developed semiphysical empirical free-energy functional

enhanced by an orientation-dependent hydrogen-bonding potential in distinguishing

native state from Rosetta docking decoys (94% success rate vs. 74% in 31

antibody-antigen and other complexes based on Z score). In addition, the monomer

potential achieved a 93% success rate in distinguishing true dimeric interfaces

from artificial crystal interfaces. More importantly, without additional

parameters, the potential provides an accurate prediction of binding free energy

of protein-peptide and protein-protein complexes (a correlation coefficient of

0.87 and a root-mean-square deviation of 1.76 kcal/mol with 69 experimental data

points). This work marks a significant step toward a unified knowledge-based

potential that quantitatively captures the common physical principle underlying

folding and binding. A Web server for academic users, established for the

prediction of binding free energy and the energy evaluation of the

protein-protein complexes, may be found at http://theory.med.buffalo.edu.

Copyright 2004 Wiley-Liss, Inc.

DOI: 10.1002/prot.20019

PMID: 15162489 [Indexed for MEDLINE]

3360. Proteins. 2004 Jul 1;56(1):11-8.

Prediction of transmembrane regions of beta-barrel proteins using ANN- and

SVM-based methods.

Natt NK(1), Kaur H, Raghava GP.

Author information:

(1)Institute of Microbial Technology, Chandigarh, India.

This article describes a method developed for predicting transmembrane

beta-barrel regions in membrane proteins using machine learning techniques:

artificial neural network (ANN) and support vector machine (SVM). The ANN used in

this study is a feed-forward neural network with a standard back-propagation

training algorithm. The accuracy of the ANN-based method improved significantly,

from 70.4% to 80.5%, when evolutionary information was added to a single sequence

as a multiple sequence alignment obtained from PSI-BLAST. We have also developed

an SVM-based method using a primary sequence as input and achieved an accuracy of

77.4%. The SVM model was modified by adding 36 physicochemical parameters to the

amino acid sequence information. Finally, ANN- and SVM-based methods were

combined to utilize the full potential of both techniques. The accuracy and

Matthews correlation coefficient (MCC) value of SVM, ANN, and combined method are

78.5%, 80.5%, and 81.8%, and 0.55, 0.63, and 0.64, respectively. These methods

were trained and tested on a nonredundant data set of 16 proteins, and

performance was evaluated using "leave one out cross-validation" (LOOCV). Based

on this study, we have developed a Web server, TBBPred, for predicting

transmembrane beta-barrel regions in proteins (available at

http://www.imtech.res.in/raghava/tbbpred).

Copyright 2004 Wiley-Liss, Inc.

DOI: 10.1002/prot.20092

PMID: 15162482 [Indexed for MEDLINE]

3361. Beijing Da Xue Xue Bao. 2004 Jun 18;36(3):322-6.

[Construction of directory for biomedical databases on INTERNET].

[Article in Chinese]

Jing X(1), Zhang QP, Guo QH, Lu M, Zhu XH, Shi L, Rui W, Shang T.

Author information:

(1)Biomedical Informatics Laboratory, Peking University Institute of

Cardiovascular Sciences, Beijing 100083, China. xiajing@bjmu.edu.cn

OBJECTIVE: To construct a global directory of biomedical databases(DBD), which

can be used free of charge on INTERNET. It will be convenient for researchers to

find out related databases quickly, easily and accurately by using DBD since

there are not enough useful tools for database retrieval. Biomedical databases

will be an accelerator in development of biomedicine with the help of DBD.

METHODS: PubMed and Google were main tools for searching related databases.

Proper search strategy with rigorous indexing rules helped us to filter

databases. The database management system was Microsoft SQL-Server 2000. The web

pages of DBD were designed with Macromedia Dreamwaver MX. ASP (active server

pages) technology was used to deal with the key words and scores sent by users.

RESULTS: There were 66 subjects and 1 258 databases in DBD at this time. We

released the Chinese and English versions of DBD on the INTERNET at the same time

http://cmbi.bjmu.edu.cn/DBList/index.htm

http://cmbi.bjmu.edu.cn/DBList/index\_en.htm . Score system was also established

to evaluate the content of the indexed databases. Users can search DBD by

subjects key words and alphabetic databases' names easily.

CONCLUSION: DBD has laid the primary foundation for further core biomedical

database evaluation system. DBD, as a useful tool for biomedical database

retrieval, will be of great aid to users since databases have played a more and

more important role in the biomedical research.

PMID: 15205710 [Indexed for MEDLINE]

3362. Bioinformatics. 2004 Jun 12;20(9):1477-9. Epub 2004 May 6.

WebSIDD: server for predicting stress-induced duplex destabilized (SIDD) sites in

superhelical DNA.

Bi C(1), Benham CJ.

Author information:

(1)UC Davis Genome Center, University of California, One Shields Avenue, Davis,

CA 95616, USA.

SUMMARY: WebSIDD is a Web-based service designed to predict locations and extents

of stress-induced duplex destabilization (SIDD) that occur in a double-stranded

DNA molecule of specified base sequence, on which a specified level of

superhelical stress is imposed. The algorithm calculates the approximate

equilibrium statistical mechanical distribution of a population of identical

molecules among its accessible states. The user inputs the DNA sequence, and the

program outputs the calculated transition probability and destabilization energy

of each base pair in the sequence. As options, the user can specify the

temperature and the level of superhelicity. The values of all structural and

energy parameters used in the calculation have been experimentally measured.

WebSIDD should prove useful for finding SIDD-susceptible sites in genomic

sequences, and correlating their occurrence with locations involved in regulatory

and pathological processes. This strategy already has illuminated the roles of

SIDD in diverse biological regulatory processes, including transcriptional

initiation and termination, and the eukaryotic nuclear scaffold attachments that

partition chromosomes into domains.

AVAILABILITY: http://orange.genomecenter.ucdavis.edu/benham/sidd/index.html

DOI: 10.1093/bioinformatics/bth304

PMID: 15130924 [Indexed for MEDLINE]

3363. Bioinformatics. 2004 Jun 12;20(9):1405-12. Epub 2004 Feb 19.

Spectral Repeat Finder (SRF): identification of repetitive sequences using

Fourier transformation.

Sharma D(1), Issac B, Raghava GP, Ramaswamy R.

Author information:

(1)Department of Biotechnology, All India Institute of Medical Sciences, New

Delhi 110029, India.

MOTIVATION: Repetitive DNA sequences, besides having a variety of regulatory

functions, are one of the principal causes of genomic instability. Understanding

their origin and evolution is of fundamental importance for genome studies. The

identification of repeats and their units helps in deducing the intra-genomic

dynamics as an important feature of comparative genomics. A major difficulty in

identification of repeats arises from the fact that the repeat units can be

either exact or imperfect, in tandem or dispersed, and of unspecified length.

RESULTS: The Spectral Repeat Finder program circumvents these problems by using a

discrete Fourier transformation to identify significant periodicities present in

a sequence. The specific regions of the sequence that contribute to a given

periodicity are located through a sliding window analysis, and an exact search

method is then used to find the repetitive units. Efficient and complete

detection of repeats is provided together with interactive and detailed

visualization of the spectral analysis of input sequence. We demonstrate the

utility of our method with various examples that contain previously unannotated

repeats. A Web server has been developed for convenient access to the automated

program.

AVAILABILITY: The Web server is available at http://www.imtech.res.in/raghava/srf

and http://www2.imtech.res.in/raghava/srf

DOI: 10.1093/bioinformatics/bth103

PMID: 14976032 [Indexed for MEDLINE]

3364. Bioinformatics. 2004 Jun 12;20(9):1468-9. Epub 2004 Feb 12.

Site2genome: locating short DNA sequences in whole genomes.

Frith MC(1), Halees AS, Hansen U, Weng Z.

Author information:

(1)Bioinformatics Program, Boston University, MA 02215, USA.

SUMMARY: Many biological papers describe short, functional DNA sites without

specifying their exact positions in the genome. We have developed a Web server

that automates the tedious task of locating such sites in eukaryotic genomes,

thus giving access to the context of rich annotations that are increasingly

available for genome sequences.

AVAILABILITY: http://zlab.bu.edu/site2genome/

DOI: 10.1093/bioinformatics/bth094

PMID: 14962939 [Indexed for MEDLINE]

3365. Bioinformatics. 2004 Jun 12;20(9):1335-60. Epub 2004 Feb 12.

Automatic prediction of protein domains from sequence information using a hybrid

learning system.

Nagarajan N(1), Yona G.

Author information:

(1)Department of Computer Science, Cornell University, Upson Hall, Ithaca, NY

14853, USA. niranjan@cs.cornell.edu

MOTIVATION: We describe a novel method for detecting the domain structure of a

protein from sequence information alone. The method is based on analyzing

multiple sequence alignments that are derived from a database search. Multiple

measures are defined to quantify the domain information content of each position

along the sequence and are combined into a single predictor using a neural

network. The output is further smoothed and post-processed using a probabilistic

model to predict the most likely transition positions between domains.

RESULTS: The method was assessed using the domain definitions in SCOP and CATH

for proteins of known structure and was compared with several other existing

methods. Our method performs well both in terms of accuracy and sensitivity. It

improves significantly over the best methods available, even some of the

semi-manual ones, while being fully automatic. Our method can also be used to

suggest and verify domain partitions based on structural data. A few examples of

predicted domain definitions and alternative partitions, as suggested by our

method, are also discussed.

AVAILABILITY: An online domain-prediction server is available at

http://biozon.org/tools/domains/

DOI: 10.1093/bioinformatics/bth086

PMID: 14962932 [Indexed for MEDLINE]

3366. BMC Bioinformatics. 2004 Jun 2;5:67.

Genome SEGE: a database for 'intronless' genes in eukaryotic genomes.

Sakharkar MK(1), Kangueane P.

Author information:

(1)Nanyang Centre for Supercomputing and Visualization, School of Mechanical and

Production Engineering, Nanyang Technological University, Singapore 639798.

mmeena@ntu.edu.sg

BACKGROUND: A number of completely sequenced eukaryotic genome data are available

in the public domain. Eukaryotic genes are either 'intron containing' or

'intronless'. Eukaryotic 'intronless' genes are interesting datasets for

comparative genomics and evolutionary studies. The SEGE database containing a

collection of eukaryotic single exon genes is available. However, SEGE is derived

using GenBank. The redundant, incomplete and heterogeneous qualities of GenBank

data are a bottleneck for biological investigation in comparative genomics and

evolutionary studies. Such studies often require representative gene sets from

each genome and this is possible only by deriving specific datasets from

completely sequenced genome data. Thus Genome SEGE, a database for 'intronless'

genes in completely sequenced eukaryotic genomes, has been constructed.

AVAILABILITY: http://sege.ntu.edu.sg/wester/intronless

DESCRIPTION: Eukaryotic 'intronless' genes are extracted from nine completely

sequenced genomes (four of which are unicellular and five of which are

multi-cellular). The complete dataset is available for download. Data subsets are

also available for 'intronless' pseudo-genes. The database provides information

on the distribution of 'intronless' genes in different genomes together with

their length distributions in each genome. Additionally, the search tool provides

pre-computed PROSITE motifs for each sequence in the database with appropriate

hyperlinks to InterPro. A search facility is also available through the web

server.

CONCLUSIONS: The unique features that distinguish Genome SEGE from SEGE is the

service providing representative 'intronless' datasets for completely sequenced

genomes. 'Intronless' gene sets available in this database will be of use for

subsequent bio-computational analysis in comparative genomics and evolutionary

studies. Such analysis may help to revisit the original genome data for

re-examination and re-annotation.

DOI: 10.1186/1471-2105-5-67

PMCID: PMC434494

PMID: 15175116 [Indexed for MEDLINE]

3367. Med Teach. 2004 Jun;26(4):336-42.

Web-based, virtual course units as a didactic concept for medical teaching.

Schultze-Mosgau S(1), Zielinski T, Lochner J.

Author information:

(1)Department of Oral and Maxillofacial Surgery, University of

Erlangen-Nuremberg, Germany. stefan.schultz-mosgau@mkg.imed.uni-erlangen.de

The objective was to develop a web-based, virtual series of lectures for

evidence-based, standardized knowledge transfer independent of location and time

with possibilities for interactive participation and a concluding web-based

online examination. Within the framework of a research project, specific Intranet

and Internet capable course modules were developed together with a concluding

examination. The concept of integrating digital and analogue course units

supported by sound was based on FlashCam (Nexus Concepts), Flash MX (Macromedia),

HTML and JavaScript. A Web server/SGI Indigo Unix server was used as a platform

by the course provider. A variety of independent formats (swf, avi, mpeg, DivX,

etc.) were integrated in the individual swf modules. An online examination was

developed to monitor the learning effect. The examination papers are

automatically forwarded by email after completion. The results are also returned

to the user automatically after they have been processed by a key program and an

evaluation program. The system requirements for the user PC have deliberately

been kept low (Internet Explorer 5.0, Flash-Player 6, 56 kbit/s modem, 200 MHz

PC). Navigation is intuitive. Users were provided with a technical online

introduction and a FAQ list. Eighty-two students of dentistry in their 3rd to 5th

years of study completed a questionnaire to assess the course content and the

user friendliness (SPSS V11) with grades 1 to 6 (1 = 'excellent' and 6 =

'unsatisfactory'). The course units can be viewed under the URL:

http://giga.rrze.uni-erlangen.de/movies/MKG/trailer and URL:

http://giga.rrze.uni-erlangen.de/movies/MKG/demo/index. Some 89% of the students

gave grades 1 (excellent) and 2 (good) for accessibility independent of time and

83% for access independent of location. Grades 1 and 2 were allocated for an

objectivization of the knowledge transfer by 67% of the students and for the use

of video sequences for demonstrating surgical techniques by 91% of the students.

The course units were used as an optional method of studying by 87% of the

students; 76% of the students made use of this facility from home; 83% of the

students used Internet Explorer as a browser; 60% used online streaming and 35%

downloading as the preferred method for data transfer. The course units

contribute to an evidence-based objectivization of multimedia knowledge transfer

independent of time and location. Online examinations permit automatic monitoring

and evaluation of the learning effect. The modular structure permits easy

updating of course contents. Hyperlinks with literature sources facilitate study.

DOI: 10.1080/01421590410001679028

PMID: 15203847 [Indexed for MEDLINE]

3368. Nihon Hoshasen Gijutsu Gakkai Zasshi. 2004 Jun;60(6):835-41.

[Development of a radiation therapy information system and linked medical image

server using techniques of WWW-DB].

[Article in Japanese]

Yokohama N(1), Kagiya G.

Author information:

(1)Medical Division, Wakasa-Wan Energy Research Center.

We developed management system for medical information such as radiation therapy

information and associated medical image information. Features of the system are

to browse medical information with web browser through network. The system was

constructed by open source software, which made proprietary client software

unnecessary. Clinical studies suggested that the system proved useful in terms of

paperless managemet. In addition, useful features are visibility and portability

to browse medical information using PDA (Personal Data Assistant) via wireless

LAN (Local Area Network). We also proposed a new approach which can contribute to

remote areas by providing medical information using the Internet.

PMID: 15220872 [Indexed for MEDLINE]

3369. Protein Eng Des Sel. 2004 Jun;17(6):527-36. Epub 2004 Aug 16.

Analysis and prediction of leucine-rich nuclear export signals.

la Cour T(1), Kiemer L, Mølgaard A, Gupta R, Skriver K, Brunak S.

Author information:

(1)Center for Biological Sequence Analysis, Biocentrum-DTU, Technical University

of Denmark, Building 208, DK-2800 Lyngby, Denmark.

We present a thorough analysis of nuclear export signals and a prediction server,

which we have made publicly available. The machine learning prediction method is

a significant improvement over the generally used consensus patterns. Nuclear

export signals (NESs) are extremely important regulators of the subcellular

location of proteins. This regulation has an impact on transcription and other

nuclear processes, which are fundamental to the viability of the cell. NESs are

studied in relation to cancer, the cell cycle, cell differentiation and other

important aspects of molecular biology. Our conclusion from this analysis is that

the most important properties of NESs are accessibility and flexibility allowing

relevant proteins to interact with the signal. Furthermore, we show that not only

the known hydrophobic residues are important in defining a nuclear export

signals. We employ both neural networks and hidden Markov models in the

prediction algorithm and verify the method on the most recently discovered NESs.

The NES predictor (NetNES) is made available for general use at

http://www.cbs.dtu.dk/.

DOI: 10.1093/protein/gzh062

PMID: 15314210 [Indexed for MEDLINE]

3370. Proteins. 2004 Jun 1;55(4):1005-13.

Single-body residue-level knowledge-based energy score combined with

sequence-profile and secondary structure information for fold recognition.

Zhou H(1), Zhou Y.

Author information:

(1)Howard Hughes Medical Institute Center for Single Molecule Biophysics,

Department of Physiology & Biophysics, State University of New York at Buffalo,

New York 14214, USA.

An elaborate knowledge-based energy function is designed for fold recognition. It

is a residue-level single-body potential so that highly efficient dynamic

programming method can be used for alignment optimization. It contains a backbone

torsion term, a buried surface term, and a contact-energy term. The energy score

combined with sequence profile and secondary structure information leads to an

algorithm called SPARKS (Sequence, secondary structure Profiles and Residue-level

Knowledge-based energy Score) for fold recognition. Compared with the popular

PSI-BLAST, SPARKS is 21% more accurate in sequence-sequence alignment in ProSup

benchmark and 10%, 25%, and 20% more sensitive in detecting the family,

superfamily, fold similarities in the Lindahl benchmark, respectively. Moreover,

it is one of the best methods for sensitivity (the number of correctly recognized

proteins), alignment accuracy (based on the MaxSub score), and specificity (the

average number of correctly recognized proteins whose scores are higher than the

first false positives) in LiveBench 7 among more than twenty servers of

non-consensus methods. The simple algorithm used in SPARKS has the potential for

further improvement. This highly efficient method can be used for fold

recognition on genomic scales. A web server is established for academic users on

http://theory.med.buffalo.edu.

Copyright 2004 Wiley-Liss, Inc.

DOI: 10.1002/prot.20007

PMID: 15146497 [Indexed for MEDLINE]

3371. Proteomics. 2004 Jun;4(6):1633-49.

Prediction of post-translational glycosylation and phosphorylation of proteins

from the amino acid sequence.

Blom N(1), Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S.

Author information:

(1)Center for Biological Sequence Analysis, The Technical University of Denmark,

Lyngby, Denmark. nikob@cbs.dtu.dk

Post-translational modifications (PTMs) occur on almost all proteins analyzed to

date. The function of a modified protein is often strongly affected by these

modifications and therefore increased knowledge about the potential PTMs of a

target protein may increase our understanding of the molecular processes in which

it takes part. High-throughput methods for the identification of PTMs are being

developed, in particular within the fields of proteomics and mass spectrometry.

However, these methods are still in their early stages, and it is indeed

advantageous to cut down on the number of experimental steps by integrating

computational approaches into the validation procedures. Many advanced methods

for the prediction of PTMs exist and many are made publicly available. We

describe our experiences with the development of prediction methods for

phosphorylation and glycosylation sites and the development of PTM-specific

databases. In addition, we discuss novel ideas for PTM visualization (exemplified

by kinase landscapes) and improvements for prediction specificity (by using

ESS--evolutionary stable sites). As an example, we present a new method for

kinase-specific prediction of phosphorylation sites, NetPhosK, which extends our

earlier and more general tool, NetPhos. The new server, NetPhosK, is made

publicly available at the URL http://www.cbs.dtu.dk/services/NetPhosK/. The

issues of underestimation, over-prediction and strategies for improving

prediction specificity are also discussed.

DOI: 10.1002/pmic.200300771

PMID: 15174133 [Indexed for MEDLINE]

3372. Bioinformatics. 2004 May 22;20(8):1322-4. Epub 2004 Feb 10.

ConSeq: the identification of functionally and structurally important residues in

protein sequences.

Berezin C(1), Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R,

Ben-Tal N.

Author information:

(1)Department of Biochemistry, Georbe S. Wise Faculty of Life Sciences, Tel Aviv

University, Israel.

MOTIVATION: ConSeq is a web server for the identification of biologically

important residues in protein sequences. Functionally important residues that

take part, e.g. in ligand binding and protein-protein interactions, are often

evolutionarily conserved and are most likely to be solvent-accessible, whereas

conserved residues within the protein core most probably have an important

structural role in maintaining the protein's fold. Thus, estimated evolutionary

rates, as well as relative solvent accessibility predictions, are assigned to

each amino acid in the sequence; both are subsequently used to indicate residues

that have potential structural or functional importance.

AVAILABILITY: The ConSeq web server is available at

http://conseq.bioinfo.tau.ac.il/

SUPPLEMENTARY INFORMATION: The ConSeq methodology, a description of its

performance in a set of five well-documented proteins, a comparison to other

methods, and the outcome of its application to a set of 111 proteins of unknown

function, are presented at http://conseq.bioinfo.tau.ac.il/ under 'OVERVIEW',

'VALIDATION', 'COMPARISON' and 'PREDICTIONS', respectively.

DOI: 10.1093/bioinformatics/bth070

PMID: 14871869 [Indexed for MEDLINE]

3373. Bioinformatics. 2004 May 22;20(8):1331-3. Epub 2004 Feb 10.

WebVar: A resource for the rapid estimation of relative site variability from

multiple sequence alignments.

Mignone F(1), Horner DS, Pesole G.

Author information:

(1)Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano,

20113 Milano, Italy.

WebVar is an online resource that provides estimates of relative site variability

from multiple alignments of homologous protein or nucleic acid sequences. WebVar

provides a variety of graphic and textual representations of estimates, designed

to assist in phylogenetic analysis.AVAILABILITY: The WebVar server is located at

http://www.pesolelab.it/Tools/WebVar.html

DOI: 10.1093/bioinformatics/bth076

PMID: 14871863 [Indexed for MEDLINE]

3374. Nucleic Acids Res. 2004 May 11;32(8):2618-22. Print 2004.

A Fugu-Human Genome Synteny Viewer: web software for graphical display and

annotation reports of synteny between Fugu genomic sequence and human genes.

Halling-Brown M(1), Sansom C, Moss DS, Elgar G, Edwards YJ.

Author information:

(1)School of Crystallography, Birkbeck College, University of London, Malet

Street, London WC1E 7HX, UK.

A web server has been developed to access annotation and graphical reports of

synteny and gene order between the Fugu genome and human genes. In this system,

the assembled Fugu genomic sequences (also known as scaffolds) are annotated. The

annotations for each Fugu scaffold are computed, stored and made publicly

available. The annotations describe matches to human homologous genes. For each

significant human gene match on the Fugu scaffold, the corresponding human

chromosome map and measures of the significance of each match are given. The

web-based server provides public access to these annotations and graphical

displays of the results. The user is provided with a selection of views including

a chromosome-colour-coded image and a table containing the details of the

matches. The Fugu-Human Genome Synteny Viewer has been tested by comparing

results with examples from a paper that includes a study of transcription

factors, Fos and Jun encoding regions. The Fugu-human genome synteny views are

available for each Fugu scaffold through the clonesearch web page located at the

Fugu Genomics website (http://fugu.rfcgr.mrc.ac.uk/).

DOI: 10.1093/nar/gkh573

PMCID: PMC419461

PMID: 15141032 [Indexed for MEDLINE]

3375. BMC Bioinformatics. 2004 May 7;5:55.

eL-DASionator: an LDAS upload file generator.

Negre V(1), Grunau C.

Author information:

(1)Institut de Génétique Humaine, CNRS UPR 1142, 141 rue de la Cardonille, 34396

Montpellier, France. developer@methdb.net

BACKGROUND: The Distributed Annotation System (DAS) allows merging of DNA

sequence annotations from multiple sources and provides a single annotation view.

A straightforward way to establish a DAS annotation server is to use the

"Lightweight DAS" server (LDAS). Onto this type of server, annotations can be

uploaded as flat text files in a defined format. The popular Ensembl ContigView

uses the same format for the transient upload and display of user data.

RESULTS: In order to easily generate LDAS upload files we developed a software

tool that is accessible via a web-interface

http://atgc.lirmm.fr/eldasionator.html. Users can submit their DNA sequences of

interest. Our program (i) aligns these sequences to the reference sequences of

Ensembl, (ii) determines start and end positions of each sequence on the

reference sequence, and (iii) generates a formatted annotation file. This file

can be used to load any LDAS annotation server or it can be uploaded to the

Ensembl ContigView.

CONCLUSION: The eL-DASionator is an on-line tool that is intended for

life-science researchers with little bioinformatics background. It conveniently

generates LDAS upload files, and makes it possible to generate annotations in a

standard format that permits comfortable sharing of this data.

DOI: 10.1186/1471-2105-5-55

PMCID: PMC416658

PMID: 15132760 [Indexed for MEDLINE]

3376. Behav Res Methods Instrum Comput. 2004 May;36(2):304-11.

Scientific LogAnalyzer: a web-based tool for analyses of server log files in

psychological research.

Reips UD(1), Stieger S.

Author information:

(1)Sozial-und Wirtchaftspsychologie, University of Zurich, Zurich, Switzerland.

u.reips@psychologie.unizh.ch

Scientific LogAnalyzer is a platform-independent interactive Web service for the

analysis of log files. Scientific LogAnalyzer offers several features not

available in other log file analysis tools--for example, organizational criteria

and computational algorithms suited to aid behavioral and social scientists.

Scientific LogAnalyzer is highly flexible on the input side (unlimited types of

log file formats), while strictly keeping a scientific output format. Features

include (1) free definition of log file format, (2) searching and marking

dependent on any combination of strings (necessary for identifying conditions in

experiment data), (3) computation of response times, (4) detection of multiple

sessions, (5) speedy analysis of large log files, (6) output in HTML and/or

tab-delimited form, suitable for import into statistics software, and (7) a

module for analyzing and visualizing drop-out. Several methodological features

specifically needed in the analysis of data collected in Internet-based

experiments have been implemented in the Web-based tool and are described in this

article. A regression analysis with data from 44 log file analyses shows that the

size of the log file and the domain name lookup are the two main factors

determining the duration of an analysis. It is less than a minute for a standard

experimental study with a 2 x 2 design, a dozen Web pages, and 48 participants

(ca. 800 lines, including data from drop-outs). The current version of Scientific

LogAnalyzer is freely available for small log files. Its Web address is

http://genpsylab-logcrunsh.unizh.ch/.

PMID: 15354696 [Indexed for MEDLINE]

3377. Bioinformatics. 2004 May 1;20(7):1157-69. Epub 2004 Feb 5.

Gene structure prediction from consensus spliced alignment of multiple ESTs

matching the same genomic locus.

Brendel V(1), Xing L, Zhu W.

Author information:

(1)Department of Genetics, Development and Cell Biology, Iowa State University,

2112 Molecular Biology Building, Ames, IA 50011-3260, USA. vbrendel@iastate.edu

MOTIVATION: Accurate gene structure annotation is a challenging computational

problem in genomics. The best results are achieved with spliced alignment of

full-length cDNAs or multiple expressed sequence tags (ESTs) with sufficient

overlap to cover the entire gene. For most species, cDNA and EST collections are

far from comprehensive. We sought to overcome this bottleneck by exploring the

possibility of using combined EST resources from fairly diverged species that

still share a common gene space. Previous spliced alignment tools were found

inadequate for this task because they rely on very high sequence similarity

between the ESTs and the genomic DNA.

RESULTS: We have developed a computer program, GeneSeqer, which is capable of

aligning thousands of ESTs with a long genomic sequence in a reasonable amount of

time. The algorithm is uniquely designed to tolerate a high percentage of

mismatches and insertions or deletions in the EST relative to the genomic

template. This feature allows use of non-cognate ESTs for gene structure

prediction, including ESTs derived from duplicated genes and homologous genes

from related species. The increased gene prediction sensitivity results in part

from novel splice site prediction models that are also available as a stand-alone

splice site prediction tool. We assessed GeneSeqer performance relative to a

standard Arabidopsis thaliana gene set and demonstrate its utility for plant

genome annotation. In particular, we propose that this method provides a timely

tool for the annotation of the rice genome, using abundant ESTs from other

cereals and plants.

AVAILABILITY: The source code is available for download at

http://bioinformatics.iastate.edu/bioinformatics2go/gs/download.html. Web servers

for Arabidopsis and other plant species are accessible at

http://www.plantgdb.org/cgi-bin/AtGeneSeqer.cgi and

http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi, respectively. For non-plant

species, use http://bioinformatics.iastate.edu/cgi-bin/gs.cgi. The splice site

prediction tool (SplicePredictor) is distributed with the GeneSeqer code. A

SplicePredictor web server is available at

http://bioinformatics.iastate.edu/cgi-bin/sp.cgi

DOI: 10.1093/bioinformatics/bth058

PMID: 14764557 [Indexed for MEDLINE]

3378. Genome Res. 2004 May;14(5):963-70.

The otter annotation system.

Searle SM(1), Gilbert J, Iyer V, Clamp M.

Author information:

(1)The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus,

Hinxton, Cambridge, CB10 1SA, UK.

With the completion of the human genome sequence and genome sequence available

for other vertebrate genomes, the task of manual annotation at the large genome

scale has become a priority. Possibly even more important, is the requirement to

curate and improve this annotation in the light of future data. For this to be

possible, there is a need for tools to access and manage the annotation. Ensembl

provides an excellent means for storing gene structures, genome features, and

sequence, but it does not support the extra textual data necessary for manual

annotation. We have extended Ensembl to create the Otter manual annotation

system. This comprises a relational database schema for storing the manual

annotation data, an application-programming interface (API) to access it, an

extensible markup language (XML) format to allow transfer of the data, and a

server to allow multiuser/multimachine access to the data. We have also written a

data-adaptor plugin for the Apollo Browser/Editor to enable it to utilize an

Otter server. The otter database is currently used by the Vertebrate Genome

Annotation (VEGA) site (http://vega.sanger.ac.uk), which provides access to

manually curated human chromosomes. Support is also being developed for using the

AceDB annotation editor, FMap, via a perl wrapper called Lace. The Human and

Vertebrate Annotation (HAVANA) group annotators at the Sanger center are using

this to annotate human chromosomes 1 and 20.

DOI: 10.1101/gr.1864804

PMCID: PMC479127

PMID: 15123593 [Indexed for MEDLINE]

3379. Hum Mutat. 2004 May;23(5):464-70.

The Swiss-Prot variant page and the ModSNP database: a resource for sequence and

structure information on human protein variants.

Yip YL(1), Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch

A.

Author information:

(1)Swiss-Prot Group, Swiss Institute of Bioinformatics, Centre Médical

Universitaire, Geneva, Switzerland. lina.yip@isb-sib.ch

Missense mutation leading to single amino acid polymorphism (SAP) is the type of

mutation most frequently related to human diseases. The Swiss-Prot protein

knowledgebase records information on such mutations in various sections of a

protein entry, namely in the "feature," "comment," and "reference" fields. To

facilitate users in obtaining the most relevant information about each human SAP

recorded in the knowledgebase, the Swiss-Prot Variant web pages were created to

provide a summary of available sequence information, as well as additional

structural information on each variant. In particular, the ModSNP database was

set up to store information related to SAPs and to manage the modeling of SAPs

onto protein structures via an automatic homology modeling pipeline. Currently,

among the 16,566 human SAPs recorded in the Swiss-Prot knowledgebase (release

42.5, 21 November 2003), more than 25% have corresponding 3D-models. Of these

variants, 47% are related to disease, 26% are polymorphisms, and 27% are not yet

clearly classified. The ModSNP database is updated and the subsequent model

construction pipeline is launched with each weekly Swiss-Prot release. Thus, the

ModSNP database represents a valuable resource for the structural analysis of

protein variation. The Swiss-Prot variant pages are accessible from the NiceProt

view of a Swiss-Prot entry on the ExPASy server (www.expasy.org/), via a

hyperlink created for the stable and unique identifier FTId of each human SAP.

Copyright 2004 Wiley-Liss, Inc.

DOI: 10.1002/humu.20021

PMID: 15108278 [Indexed for MEDLINE]

3380. Virchows Arch. 2004 May;444(5):403-9. Epub 2004 Mar 12.

Diagnostic telepathology: long-term experience of a single institution.

Brauchli K(1), Oberli H, Hurwitz N, Kunze KD, Haroske G, Jundt G, Stauch G,

Banach L, Wirdnam M, Mihatsch M, Oberholzer M.

Author information:

(1)Department of Pathology of the University, Basel, Switzerland.

OBJECTIVES: The paper reviews the development of the application of telepathology

in a department of surgical pathology between 1991 and 2003. The goal of the

efforts during this time was to give up the concept of programming a single

application, available only between two fixed workstations with sophisticated

devices and special software, and to find the virtual "largest common

denominator" for implementing as many different applications as possible with the

same basic system.

METHODS: A new telepathology system was designed as a client-server system with a

relational database at its centre. The clients interact together by transferring

the questions (texts and images) to a record (case) in the database on the server

and by transferring the answers to the same record on the database.

RESULTS: The new "open" telepathology system iPath

(http://telepath.patho.unibas.ch) has been very well accepted by many groups

around the world. The main application fields are: consultations between

pathologists and medical institutions without a pathologist (e.g. for frozen

section diagnoses or for surgical diagnoses in hospitals in South Asia or

Africa), tumour boards, field studies and distance education

(http://teleteach.patho.unibas.ch).

CONCLUSIONS: Having observed that with iPath we have succeeded in satisfying all

our telepathology needs, we are inclined to put the emphasis on the nature of the

tasks being performed, as opposed to the methods or technical means for

performing a given task. The three organisation models proposed by Weinstein et

al. (2001) can be reduced to only two models: the model of discussion groups and

the model of expert groups (virtual institutes).

DOI: 10.1007/s00428-004-0980-x

PMID: 15021986 [Indexed for MEDLINE]

3381. J Comput Chem. 2004 Apr 15;25(5):762-7.

Neural network-based prediction of transmembrane beta-strand segments in outer

membrane proteins.

Gromiha MM(1), Ahmad S, Suwa M.

Author information:

(1)Computational Biology Research Center, National Institute of Advanced

Industrial Science and Technology, Aomi Frontier Building 17F, 2-43 Aomi,

Koto-ku, Tokyo 135-0064, Japan. michael-gromiha@aist.go.jp

Prediction of transmembrane beta-strands in outer membrane proteins (OMP) is one

of the important problems in computational chemistry and biology. In this work,

we propose a method based on neural networks for identifying the

membrane-spanning beta-strands. We introduce the concept of "residue probability"

for assigning residues in transmembrane beta-strand segments. The performance of

our method is evaluated with single-residue accuracy, correlation, specificity,

and sensitivity. Our predicted segments show a good agreement with experimental

observations with an accuracy level of 73% solely from amino acid sequence

information. Further, the predictive power of N- and C-terminal residues in each

segments, number of segments in each protein, and the influence of cutoff

probability for identifying membrane-spanning beta-strands will be discussed. We

have developed a Web server for predicting the transmembrane beta-strands from

the amino acid sequence, and the prediction results are available at

http://psfs.cbrc.jp/tmbeta-net/.

Copyright 2004 Wiley Periodicals, Inc. J Comput Chem 25: 762-767, 2004

DOI: 10.1002/jcc.10386

PMID: 14978719 [Indexed for MEDLINE]

3382. J Thorac Imaging. 2004 Apr;19(2):103-8.

Casimage project: a digital teaching files authoring environment.

Rosset A(1), Muller H, Martins M, Dfouni N, Vallée JP, Ratib O.

Author information:

(1)UCLA Radiological Sciences Department, Los Angeles, California 90095-1721,

USA.

The goal of the Casimage project is to offer an authoring and editing environment

integrated with the Picture Archiving and Communication Systems (PACS) for

creating image-based electronic teaching files. This software is based on a

client/server architecture allowing remote access of users to a central database.

This authoring environment allows radiologists to create reference databases and

collection of digital images for teaching and research directly from clinical

cases being reviewed on PACS diagnostic workstations. The environment includes

all tools to create teaching files, including textual description, annotations,

and image manipulation. The software also allows users to generate stand-alone

CD-ROMs and web-based teaching files to easily share their collections. The

system includes a web server compatible with the Medical Imaging Resource Center

standard (MIRC, http://mirc.rsna.org) to easily integrate collections in the RSNA

web network dedicated to teaching files. This software could be installed on any

PACS workstation to allow users to add new cases at any time and anywhere during

clinical operations. Several images collections were created with this tool,

including thoracic imaging that was subsequently made available on a CD-Rom and

on our web site and through the MIRC network for public access.

PMID: 15071328 [Indexed for MEDLINE]

3383. Proteins. 2004 Apr 1;55(1):83-90.

Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary

structure information.

Kaur H(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Chandigarh, India.

In this paper a systematic attempt has been made to develop a better method for

predicting alpha-turns in proteins. Most of the commonly used approaches in the

field of protein structure prediction have been tried in this study, which

includes statistical approach "Sequence Coupled Model" and machine learning

approaches; i) artificial neural network (ANN); ii) Weka (Waikato Environment for

Knowledge Analysis) Classifiers and iii) Parallel Exemplar Based Learning

(PEBLS). We have also used multiple sequence alignment obtained from PSIBLAST and

secondary structure information predicted by PSIPRED. The training and testing of

all methods has been performed on a data set of 193 non-homologous protein X-ray

structures using five-fold cross-validation. It has been observed that ANN with

multiple sequence alignment and predicted secondary structure information

outperforms other methods. Based on our observations we have developed an

ANN-based method for predicting alpha-turns in proteins. The main components of

the method are two feed-forward back-propagation networks with a single hidden

layer. The first sequence-structure network is trained with the multiple sequence

alignment in the form of PSI-BLAST-generated position specific scoring matrices.

The initial predictions obtained from the first network and PSIPRED predicted

secondary structure are used as input to the second structure-structure network

to refine the predictions obtained from the first net. The final network yields

an overall prediction accuracy of 78.0% and MCC of 0.16. A web server AlphaPred

(http://www.imtech.res.in/raghava/alphapred/) has been developed based on this

approach.

Copyright 2004 Wiley-Liss, Inc.

DOI: 10.1002/prot.10569

PMID: 14997542 [Indexed for MEDLINE]

3384. Bioinformatics. 2004 Mar 22;20(5):727-34. Epub 2004 Jan 29.

Java-based application framework for visualization of gene regulatory region

annotations.

Sun H(1), Davuluri RV.

Author information:

(1)Human Cancer Genetics Program, Comprehensive Cancer Center, Department of

Molecular Virology, Immunology and Medical Genetics, The Ohio State University,

420 W 12th Avenue, TMRF 524, Columbus, OH 43210, USA.

MOTIVATION: The genome sequences of several organisms are either complete, or

being sequenced. Each genome needs to be integrated with various types of

annotations, e.g. locations of genes, promoters and other functional elements

such as transcriptional regulatory elements. A robust application framework will

be useful for developing web-based applications to visualize various genome

annotations.

RESULTS: We developed genome data visualization toolkit (GDVTK) as an application

framework that consists of a set of data structures and core classes, using Java

technology. GDVTK is a sound framework for developing web-based applications to

present the gene regulatory region annotations in visual form. The current

version of GDVTK consists of eight packages and 38 Java classes that are

portable, reusable and extensible for plugging in new data sources and models. We

implemented GDVTK for visualization of promoter annotations in Mammalian Promoter

Database (MPromDb), a web-based gene-regulatory information server.

AVAILABILITY: GDVTK is available under GNU general public license. Source code

and software documentation can be found at the URL

http://bioinformatics.med.ohio-state.edu/GDVTK.

DOI: 10.1093/bioinformatics/btg476

PMID: 14751988 [Indexed for MEDLINE]

3385. Bioinformatics. 2004 Mar 22;20(5):805-7. Epub 2004 Jan 29.

ClusterControl: a web interface for distributing and monitoring bioinformatics

applications on a Linux cluster.

Stocker G(1), Rieder D, Trajanoski Z.

Author information:

(1)Institute of Biomedical Engineering and Christian-Doppler-Laboratory for

Genomics and Bioinformatics, Graz University of Technology, Krenngasse 37, 8010

Graz, Austria.

ClusterControl is a web interface to simplify distributing and monitoring

bioinformatics applications on Linux cluster systems. We have developed a modular

concept that enables integration of command line oriented program into the

application framework of ClusterControl. The systems facilitate integration of

different applications accessed through one interface and executed on a

distributed cluster system. The package is based on freely available technologies

like Apache as web server, PHP as server-side scripting language and OpenPBS as

queuing system and is available free of charge for academic and non-profit

institutions.AVAILABILITY: http://genome.tugraz.at/Software/ClusterControl

DOI: 10.1093/bioinformatics/bth014

PMID: 14751976 [Indexed for MEDLINE]

3386. J Mol Biol. 2004 Mar 19;337(2):243-53.

A sensitive predictor for potential GPI lipid modification sites in fungal

protein sequences and its application to genome-wide studies for Aspergillus

nidulans, Candida albicans, Neurospora crassa, Saccharomyces cerevisiae and

Schizosaccharomyces pombe.

Eisenhaber B(1), Schneider G, Wildpaner M, Eisenhaber F.

Author information:

(1)Research Institute of Molecular Pathology (IMP), Dr Bohr-Gasse 7, A-1030

Vienna, Austria. eisenhaber@imp.univie.ac.at

The fungal transamidase complex that executes glycosylphosphatidylinositol (GPI)

lipid anchoring of precursor proteins has overlapping but distinct sequence

specificity compared with the animal system. Therefore, a taxon-specific

prediction tool for the recognition of the C-terminal signal in fungal sequences

is necessary. We have collected a learning set of fungal precursor protein

sequences from the literature and fungal proteomes. Although the general four

segment scheme of the recognition signal is maintained also in fungal precursors,

there are taxon specificities in details. A fungal big-Pi predictor has been

developed for the assessment of query sequence concordance with fungi-specific

recognition signal requirements. The sensitivity of this predictor is close to

90%. The rate of false positive prediction is in the range of 0.1%. The fungal

big-Pi tool successfully predicts the Gas1 mutation series described by C.

Nuoffer and co-workers, and recognizes that the human PLAP C terminus is not a

target for the fungal transamidase complex. Lists of potentially GPI lipid

anchored proteins for five fungal proteomes have been generated and the hits have

been functionally classified. The fungal big-Pi prediction WWW server as well as

precursor lists are available at

DOI: 10.1016/j.jmb.2004.01.025

PMID: 15003443 [Indexed for MEDLINE]

3387. Med Inform Internet Med. 2004 Mar;29(1):75-85.

An online tool for investigating clinical decision making.

Parry DT(1), Parry EC, Pattison NS.

Author information:

(1)School of Computer and Information Sciences, Auckland University of

Technology, New Zealand. dave.parry@aut.ac.nz

BACKGROUND: Induction of labour is a common clinical intervention. There has been

a recent rise in rates of induction of labour and wide variation between

published hospital rates without obvious explanation. Clinician variation has

been suggested as a reason.

OBJECTIVE: The study described aimed to examine clinical decision making, whilst

removing individual patient bias. To achieve this clinical behaviour was studied

by the use of imaginary clinical scenarios presented to clinicians by computer.

Unlike retrospective audit, the rates thus generated are unaffected by

differences in casemix, pressure of time, work or other factors and allow direct

comparison between clinicians and comparison with clinical guidelines.

METHODS: Data about 15 imaginary pregnant women are presented to the clinician,

each may have symptoms or signs of hypertensive disorders, intrauterine growth

restriction (IUGR) and/or postdates. From the decision made in each scenario, and

the information revealed about each scenario, a set of 'decision rules' is

created for each clinician, describing in what circumstances they would induce

labour. Data from the National Women's Hospital (Auckland, New Zealand) is then

examined using these rules and the induction of labour rate thus generated

presented to the clinician.

RESULTS: Sixteen clinicians were interviewed. Their induction of labour rate

ranged from 10-31%.

CONCLUSIONS: Clinician variation in decision making is evident about the

intervention when to induce labour. The system is available on the WWW at

http://csrs2.aut.ac.nz/scenario

DOI: 10.1080/14639230410001662660

PMID: 15204612 [Indexed for MEDLINE]

3388. Proteins. 2004 Mar 1;54(4):738-43.

A novel method for protein secondary structure prediction using dual-layer SVM

and profiles.

Guo J(1), Chen H, Sun Z, Lin Y.

Author information:

(1)Institute of Bioinformatics, State Key Laboratory of Biomembrane and Membrane

Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua

University, Beijing, China.

A high-performance method was developed for protein secondary structure

prediction based on the dual-layer support vector machine (SVM) and

position-specific scoring matrices (PSSMs). SVM is a new machine learning

technology that has been successfully applied in solving problems in the field of

bioinformatics. The SVM's performance is usually better than that of traditional

machine learning approaches. The performance was further improved by combining

PSSM profiles with the SVM analysis. The PSSMs were generated from PSI-BLAST

profiles, which contain important evolution information. The final prediction

results were generated from the second SVM layer output. On the CB513 data set,

the three-state overall per-residue accuracy, Q3, reached 75.2%, while segment

overlap (SOV) accuracy increased to 80.0%. On the CB396 data set, the Q3 of our

method reached 74.0% and the SOV reached 78.1%. A web server utilizing the method

has been constructed and is available at

http://www.bioinfo.tsinghua.edu.cn/pmsvm.

Copyright 2004 Wiley-Liss, Inc.

DOI: 10.1002/prot.10634

PMID: 14997569 [Indexed for MEDLINE]

3389. Protein Sci. 2004 Feb;13(2):391-9.

Accurate and efficient loop selections by the DFIRE-based all-atom statistical

potential.

Zhang C(1), Liu S, Zhou Y.

Author information:

(1)Howard Hughes Medical Institute Center for Single Molecule Biophysics and

Department of Physiology and Biophysics, State University of New York at Buffalo,

124 Sherman Hall, Buffalo, NY 14214, USA.

The conformations of loops are determined by the water-mediated interactions

between amino acid residues. Energy functions that describe the interactions can

be derived either from physical principles (physical-based energy function) or

statistical analysis of known protein structures (knowledge-based statistical

potentials). It is commonly believed that statistical potentials are appropriate

for coarse-grained representation of proteins but are not as accurate as

physical-based potentials when atomic resolution is required. Several recent

applications of physical-based energy functions to loop selections appear to

support this view. In this article, we apply a recently developed DFIRE-based

statistical potential to three different loop decoy sets (RAPPER, Jacobson, and

Forrest-Woolf sets). Together with a rotamer library for side-chain optimization,

the performance of DFIRE-based potential in the RAPPER decoy set (385 loop

targets) is comparable to that of AMBER/GBSA for short loops (two to eight

residues). The DFIRE is more accurate for longer loops (9 to 12 residues).

Similar trend is observed when comparing DFIRE with another physical-based

OPLS/SGB-NP energy function in the large Jacobson decoy set (788 loop targets).

In the Forrest-Woolf decoy set for the loops of membrane proteins, the DFIRE

potential performs substantially better than the combination of the CHARMM force

field with several solvation models. The results suggest that a single-term

DFIRE-statistical energy function can provide an accurate loop prediction at a

fraction of computing cost required for more complicate physical-based energy

functions. A Web server for academic users is established for loop selection at

the softwares/services section of the Web site http://theory.med.buffalo.edu/.

DOI: 10.1110/ps.03411904

PMCID: PMC2286705

PMID: 14739324 [Indexed for MEDLINE]

3390. Bioinformatics. 2004 Jan 22;20(2):282-3.

FGDP: functional genomics data pipeline for automated, multiple microarray data

analyses.

Grant JD(1), Somers LA, Zhang Y, Manion FJ, Bidaut G, Ochs MF.

Author information:

(1)Bioinformatics, Department of Information Science and Technology, Division of

Basic Science, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, PA

19111, USA.

Gene expression microarrays and oligonucleotide GeneChips have provided

biologists with a means of measuring, in a single experiment, the expression

levels of entire genomes under a variety of conditions. As with any nascent

field, there is no single accepted method for analyzing the new data types, with

new methods appearing monthly. Investigators using the new technology must

constantly seek access to the latest tools and explore their data in multiple

ways. The functional genomics data pipeline provides an integrated, extendable

analysis environment permitting multiple, simultaneous analyses to be

automatically performed and provides a web server and interface for presenting

results.AVAILABILITY: Source code and executables are available under the GNU

public license at http://bioinformatics.fccc.edu/

PMID: 14734324 [Indexed for MEDLINE]

3391. Bioinformatics. 2004 Jan 22;20(2):279-81.

JDotter: a Java interface to multiple dotplots generated by dotter.

Brodie R(1), Roper RL, Upton C.

Author information:

(1)Department of Biochemistry and Microbiology, University of Victoria, Victoria,

BC V8W 3P6, Canada.

Java-Dotter (JDotter) is a platform-independent Java interactive interface for

the Linux version of Dotter, a widely used program for generating dotplots of

large DNA or protein sequences. JDotter runs as a client-server application and

can send new sequences to the Dotter program for alignment as well as rapidly

access a repository of preprocessed dotplots. JDotter also interfaces with a

sequence database or file system to display supplementary feature data. Thus,

JDotter greatly simplifies access to dotplot data in laboratories that deal with

large numbers of genomes and have a multi-platform organization.AVAILABILITY:

Currently, JDotter is used via Java Web Start by the Poxvirus Bioinformatics

Resource for examining dotplots of complete poxvirus genomes;

http://athena.bioc.uvic.ca/pbr/jdotter/. The software is available for download

from the same location.

SUPPLEMENTARY INFORMATION: Installation instructions, the User's Manual,

screenshots and examples are available at the JDotter home page

http://athena.bioc.uvic.ca/pbr/jdotter/. The software and source code is free for

non-commercial applications.

PMID: 14734323 [Indexed for MEDLINE]

3392. BMC Bioinformatics. 2004 Jan 21;5:7.

HMM Logos for visualization of protein families.

Schuster-Böckler B(1), Schultz J, Rahmann S.

Author information:

(1)Department of Mathematics and Computer Science, Freie Universität Berlin,

Germany. bendaboots@gmx.de

BACKGROUND: Profile Hidden Markov Models (pHMMs) are a widely used tool for

protein family research. Up to now, however, there exists no method to visualize

all of their central aspects graphically in an intuitively understandable way.

RESULTS: We present a visualization method that incorporates both emission and

transition probabilities of the pHMM, thus extending sequence logos introduced by

Schneider and Stephens. For each emitting state of the pHMM, we display a stack

of letters. The stack height is determined by the deviation of the position's

letter emission frequencies from the background frequencies. The stack width

visualizes both the probability of reaching the state (the hitting probability)

and the expected number of letters the state emits during a pass through the

model (the state's expected contribution).A web interface offering online

creation of HMM Logos and the corresponding source code can be found at the Logos

web server of the Max Planck Institute for Molecular Genetics

http://logos.molgen.mpg.de.

CONCLUSIONS: We demonstrate that HMM Logos can be a useful tool for the

biologist: We use them to highlight differences between two homologous

subfamilies of GTPases, Rab and Ras, and we show that they are able to indicate

structural elements of Ras.

DOI: 10.1186/1471-2105-5-7

PMCID: PMC341448

PMID: 14736340 [Indexed for MEDLINE]

3393. BMC Bioinformatics. 2004 Jan 9;5:2.

Using 3D Hidden Markov Models that explicitly represent spatial coordinates to

model and compare protein structures.

Alexandrov V(1), Gerstein M.

Author information:

(1)Department of Molecular Biophysics and Biochemistry, Yale University, 266

Whitney Ave, New Haven, CT 06511, USA. vadim.alexandrov@yale.edu

BACKGROUND: Hidden Markov Models (HMMs) have proven very useful in computational

biology for such applications as sequence pattern matching, gene-finding, and

structure prediction. Thus far, however, they have been confined to representing

1D sequence (or the aspects of structure that could be represented by character

strings).

RESULTS: We develop an HMM formalism that explicitly uses 3D coordinates in its

match states. The match states are modeled by 3D Gaussian distributions centered

on the mean coordinate position of each alpha carbon in a large structural

alignment. The transition probabilities depend on the spread of the neighboring

match states and on the number of gaps found in the structural alignment. We also

develop methods for aligning query structures against 3D HMMs and scoring the

result probabilistically. For 1D HMMs these tasks are accomplished by the Viterbi

and forward algorithms. However, these will not work in unmodified form for the

3D problem, due to non-local quality of structural alignment, so we develop

extensions of these algorithms for the 3D case. Several applications of 3D HMMs

for protein structure classification are reported. A good separation of scores

for different fold families suggests that the described construct is quite useful

for protein structure analysis.

CONCLUSION: We have created a rigorous 3D HMM representation for protein

structures and implemented a complete set of routines for building 3D HMMs in C

and Perl. The code is freely available from

http://www.molmovdb.org/geometry/3dHMM, and at this site we also have a simple

prototype server to demonstrate the features of the described approach.

DOI: 10.1186/1471-2105-5-2

PMCID: PMC344530

PMID: 14715091 [Indexed for MEDLINE]

3394. Acta Biochim Pol. 2004;51(1):161-72.

Comparison of proteins based on segments structural similarity.

Plewczynski D(1), Pas J, Von Grotthuss M, Rychlewski L.

Author information:

(1)Interdisciplinary Center for Mathematical and Computational Modeling Warsaw

University, Warszawa, Poland. darman@bioinfo.pl

We present here a simple method for fast and accurate comparison of proteins

using their structures. The algorithm is based on structural alignment of

segments of Calpha chains (with size of 99 or 199 residues). The method is

optimized in terms of speed and accuracy. We test it on 97 representative

proteins with the similarity measure based on the SCOP classification. We compare

our algorithm with the LGscore2 automatic method. Our method has the same

accuracy as the LGscore2 algorithm with much faster processing of the whole test

set, which is promising. A second test is done using the ToolShop structure

prediction evaluation program and shows that our tool is on average slightly less

sensitive than the DALI server. Both algorithms give a similar number of correct

models, however, the final alignment quality is better in the case of DALI. Our

method was implemented under the name 3D-Hit as a web server at

http://3dhit.bioinfo.pl/ free for academic use, with a weekly updated database

containing a set of 5000 structures from the Protein Data Bank with

non-homologous sequences.

DOI: 045101161

PMID: 15094837 [Indexed for MEDLINE]

3395. Appl Bioinformatics. 2004;3(4):253-6.

Biosphere: the interoperation of web services in microarray cluster analysis.

Cheung KH(1), de Knikker R, Guo Y, Zhong G, Hager J, Yip KY, Kwan AK, Li P,

Cheung DW.

Author information:

(1)Center for Medical Informatics, Yale University, 300 George Street, New Haven,

CT 06511, USA. kei.cheung@yale.edu

The growing use of DNA microarrays in biomedical research has led to the

proliferation of analysis tools. These software programs address different

aspects of analysis (e.g. normalisation and clustering within and across

individual arrays) as well as extended analysis methods (e.g. clustering,

annotation and mining of multiple datasets). Therefore, microarray data analysis

typically requires the interoperability of multiple software programs involving

different analysis types and methods. Such interoperation is often hampered by

the heterogeneity inherent in the software tools (which may function by

implementing different interfaces and using different programming languages). To

address this problem, we employed the simple object access protocol (SOAP)-based

web service approach that provides a uniform programmatic interface to these

heterogeneous software components. To demonstrate this approach in the microarray

context, we created a web server application, Biosphere, which interoperates a

number of web services that are geographically widely distributed. These web

services include a clustering web service, which is a suite of different

clustering algorithms for analysing microarray data; XEMBL, developed at the

European Bioinformatics Institute (EBI) for retrieving EMBL Nucleotide Sequence

Database sequence data; and three gene annotation web services: GetGO, GetHAPI

and GetUMLS. GetGO allows retrieval of Gene Ontology (GO) annotation, and the

other two web services retrieve annotation from the biomedical literature that is

indexed based on the Medical Subject Headings (MeSH) terms. With these web

services, Biosphere allows the users to do the following: (i) cluster gene

expression data using seven different algorithms; (ii) visualise the clustering

results that are grouped statistically in colour; and (iii) retrieve sequence,

annotation and citation data for the genes of interest.AVAILABILITY: Biosphere

and its web services described in Web Service Description Language (WSDL) can be

accessed at http://rook.cecid.hku.hk:8280/BiosphereServer.

PMID: 15702956 [Indexed for MEDLINE]

3396. Appl Bioinformatics. 2004;3(4):237-40.

HCVDB: hepatitis C virus sequences database.

Combet C(1), Penin F, Geourjon C, Deléage G.

Author information:

(1)Institut de Biologie et Chimie des Protéines, UMR 5086, CNRS/UCBL, IFR128

Biosciences Lyon-Gerland, 7 passage du Vercors, 69367 Lyon Cedex 07, France.

To date, more than 30 000 hepatitis C virus (HCV) sequences have been deposited

in the generalist databases DNA Data Bank of Japan (DDBJ), EMBL Nucleotide

Sequence Database (EMBL) and GenBank. The main difficulties with HCV sequences in

these databases are their retrieval, annotation and analyses. To help HCV

researchers face the increasing needs of HCV sequence analyses, we developed a

specialised database of computer-annotated HCV sequences, called HCVDB. HCVDB is

re-built every month from an up-to-date EMBL database by an automated process.

HCVDB provides key data about the HCV sequences (e.g. genotype, genomic region,

protein names and functions, known 3-dimensional structures) and ensures

consistency of the annotations, which enables reliable keyword queries. The

database is highly integrated with sequence and structure analysis tools and the

SRS (LION bioscience) keywords query system. Thus, any user can extract subsets

of sequences matching particular criteria or enter their own sequences and

analyse them with various bioinformatics programs available on the same

server.AVAILABILITY: HCVDB is available from http://hepatitis.ibcp.fr.

PMID: 15702954 [Indexed for MEDLINE]

3397. Biofactors. 2004;22(1-4):329-32.

The construction of web database server-client system for functional food

factors.

Zhuo XG(1), Watanabe S.

Author information:

(1)Study Group on Functional Food Factors, Department of Nutrition Science,

Faculty of Applied Bioscience, Tokyo University of Agriculture, 1-1-1 Sakuragaoka

Setagaya, Tokyo 156-8502, Japan.

In food, other than known nutrients, such as lipid, carbohydrate, protein,

vitamins, and minerals, many substances with physiological function and medicinal

action exist, and it is contributing to healthy improvement and/or prevention of

illness. Although carotenoid, flavonoid and polyphenol, terpenoid, volatile

substance and sulfur compounds, peptide, etc. have the function of illness

prevention, and research of those non-nutrient functional food factors (FFF)

became globally active, the research of this field is not yet done

systematically. We evaluate function of FFF and reappraise known knowledge, and

this knowledge is standardized and accumulated, aimed at building a web database

server-client system which is easy to use for the people and nutritional

research. We also collected related data such as chemical characters of FFF from

literatures and other source, and formatted them into the database. We

constructed the web database server-client system with MySQL database server and

Apache web server based on Linux, and used Tomcat JSP engine for data connecting

since they were reliable in stability and speed. We are opening the database at

http://www.life-science.jp/FFF for test now.

PMID: 15630306 [Indexed for MEDLINE]

3398. Bioinformatics. 2004 Jan 1;20(1):136-7.

TM or not TM: transmembrane protein prediction with low false positive rate using

DAS-TMfilter.

Cserzo M(1), Eisenhaber F, Eisenhaber B, Simon I.

Author information:

(1)Agricultural Biotechnology Center, Szent-Gyorgyi A. 4, Gödöllö, H-2100

Hungary.

Web-based servers implementing the DAS-TMfilter algorithm have been launched at

three mirror sites and their usage is described. The underlying computer program

is an upgraded and modified version of the DAS-prediction method. The new server

is (approximately 1 among 100 unrelated queries) while the high efficiency of the

original algorithm locating TM segments in queries is preserved (sensitivity of

approximately 95% among documented proteins with helical TM

regions).AVAILABILITY: The server operates at three mirror sites:

http://mendel.imp.univie.ac.at/sat/DAS/DAS.html,

http://wooster.bip.bham.ac.uk/DAS.html and http://www.enzim.hu/DAS/DAS.html. The

program is available on request.

PMID: 14693825 [Indexed for MEDLINE]

3399. Bioinformatics. 2004 Jan 1;20(1):75-7.

MedBlast: searching articles related to a biological sequence.

Tu Q(1), Tang H, Ding D.

Author information:

(1)Key Laboratory of Proteomics, Institute of Biochemistry and Cell Biology,

Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences,

Shanghai 200031, People's Republic of China.

In the genomic era, researchers often want to know more information about a

biological sequence by retrieving its related articles. However, there is no

available tool yet to achieve conveniently this goal. Here we developed a new

literature-mining tool MedBlast, which uses natural language processing

techniques, to retrieve the related articles of a given sequence. An online

server of this program is also provided.AVAILABILITY: Both online server and the

program are available freely at http://medblast.sibsnet.org

PMID: 14693811 [Indexed for MEDLINE]

3400. Bioinformatics. 2004 Jan 1;20(1):45-50.

ClusPro: an automated docking and discrimination method for the prediction of

protein complexes.

Comeau SR(1), Gatchell DW, Vajda S, Camacho CJ.

Author information:

(1)Bioinformatics Graduate Program and Department, Boston University, 44

Cummington St, Boston, MA 02215, USA.

MOTIVATION: Predicting protein interactions is one of the most challenging

problems in functional genomics. Given two proteins known to interact, current

docking methods evaluate billions of docked conformations by simple scoring

functions, and in addition to near-native structures yield many false positives,

i.e. structures with good surface complementarity but far from the native.

RESULTS: We have developed a fast algorithm for filtering docked conformations

with good surface complementarity, and ranking them based on their clustering

properties. The free energy filters select complexes with lowest desolvation and

electrostatic energies. Clustering is then used to smooth the local minima and to

select the ones with the broadest energy wells-a property associated with the

free energy at the binding site. The robustness of the method was tested on sets

of 2000 docked conformations generated for 48 pairs of interacting proteins. In

31 of these cases, the top 10 predictions include at least one near-native

complex, with an average RMSD of 5 A from the native structure. The docking and

discrimination method also provides good results for a number of complexes that

were used as targets in the Critical Assessment of PRedictions of Interactions

experiment.

AVAILABILITY: The fully automated docking and discrimination server ClusPro can

be found at http://structure.bu.edu

PMID: 14693807 [Indexed for MEDLINE]

3401. Conf Proc IEEE Eng Med Biol Soc. 2004;3:2200-3.

A PDA-based flexible telecommunication system for telemedicine applications.

Nazeran H(1), Setty S, Haltiwanger E, Gonzalez V.

Author information:

(1)Department of Electrical and Computer Engineering, University of Texas, El

Paso, TX 79968, USA. nazeran@ece.utep.edu

Technology has been used to deliver health care at a distance for many years.

Telemedicine is a rapidly growing area and recently there are studies devoted to

prehospital care of patients in emergency cases. In this work we have developed a

compact, reliable, and low cost PDA-based telecommunication device for

telemedicine applications to transmit audio, still images, and vital signs from a

remote site to a fixed station such as a clinic or a hospital in real time. This

was achieved based on a client-server architecture. A Pocket PC, a miniature

camera, and a hands-free microphone were used at the client site and a desktop

computer running the Windows XP operating system was used as a server. The server

was located at a fixed station. The system was implemented on TCP/IP and HTTP

protocol. Field tests have shown that the system can reliably transmit still

images, audio, and sample vital signs from a simulated remote site to a fixed

station either via a wired or wireless network in real time. The Pocket PC was

used at the client site because of its compact size, low cost and processing

capabilities.

DOI: 10.1109/IEMBS.2004.1403642

PMID: 17272162

3402. In Silico Biol. 2004;4(4):411-5.

A search tool for identification and analysis of conserved sequence patterns in

Saccharomyces spp. orthologous promoter.

Kohli DK(1), Srikanth CV, Bachhawat AK.

Author information:

(1)Institute of Microbial Technology, Sector 39-A, Chandigarh-160 036, India.

We describe a web-based resource to identify, search and analyze sequence

patterns conserved in the multiple sequence alignments of orthologous promoters

from closely related / distant Saccharomyces spp. The webtool interfaces with a

database where conserved sequence patterns (greater than 4 bp) have been

previously extracted from genome-wide promoter alignments, allowing one to carry

out user-defined genome-wide searches for conserved sequences to assist in the

discovery of novel promoter elements based on comparative genomics. The web-based

server can be accessed at http://www2.imtech.res.in/ anand/sacch\_prom\_pat.html.

PMID: 15506991 [Indexed for MEDLINE]

3403. In Silico Biol. 2004;4(2):127-31.

waveTM: wavelet-based transmembrane segment prediction.

Pashou EE(1), Litou ZI, Liakopoulos TD, Hamodrakas SJ.

Author information:

(1)Department of Cell Biology and Biophysics, Faculty of Biology, University of

Athens, Panepistimiopolis, Athens 157 01, Greece.

waveTM is a web tool for the prediction of transmembrane segments in

alpha-helical membrane proteins. Prediction is performed by a dynamic programming

algorithm on wavelet-denoised 'hydropathy' signals. Users submit a protein

sequence and receive interactively the results. Topology prediction can also be

obtained in conjunction with the algorithm OrienTM. A web server that implements

the waveTM algorithm is freely available at

http://bioinformatics.biol.uoa.gr/waveTM.

PMID: 15107018 [Indexed for MEDLINE]

3404. J Mol Graph Model. 2004 Jan;22(3):195-207.

Quantitative online prediction of peptide binding to the major histocompatibility

complex.

Hattotuwagama CK(1), Guan P, Doytchinova IA, Zygouri C, Flower DR.

Author information:

(1)Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG20 7NN, UK.

With its implications for vaccine discovery, the accurate prediction of T cell

epitopes is one of the key aspirations of computational vaccinology. We have

developed a robust multivariate statistical method, based on partial least

squares, for the quantitative prediction of peptide binding to major

histocompatibility complexes (MHC), the principal checkpoint on the antigen

presentation pathway. As a service to the immunobiology community, we have made a

Perl implementation of the method available via a World Wide Web server. We call

this server MHCPred. Access to the server is freely available from the URL:

http://www.jenner.ac.uk/MHCPred. We have exemplified our method with a model for

peptides binding to the common human MHC molecule HLA-B\*3501.

DOI: 10.1016/S1093-3263(03)00160-8

PMID: 14629978 [Indexed for MEDLINE]

3405. J Struct Funct Genomics. 2004;5(1-2):13-21.

Primer Prim'er: a web based server for automated primer design.

Everett JK(1), Acton TB, Montelione GT.

Author information:

(1)Department of Biochemistry, Robert Wood Johnson Medical School, Piscataway, NJ

08854-5638, USA.

The Northeast Structural Genomics Consortium (NESG) is one of nine NIH-funded

pilot projects created to develop technologies needed for structural studies of

proteins on a genome-wide scale. One of the most challenging aspects of this

emerging field is the production of protein samples amenable to structural

determination. To do this efficiently, all steps in the protein production

pipeline must be automated. Here we describe the Primer program (linked from

http://www-nmr.cabm.rutgers.edu/bioinformatics,

www-nmr.cabm.rutgers.edu/bioinformatics, a web-based primer design program freely

available to the scientific community, which was created to automate this time

consuming and laborious task. This program has the ability to simultaneously

calculate plasmid specific primer sets for multiple open reading frame (ORF)

targets, including 96-well and greater formats. Primer includes a library of

commonly used plasmid systems and possesses the ability to upload user-defined

plasmid systems. In addition to calculating gene-specific annealing regions for

each target, the program also adds appropriate restriction endonuclease

recognition or viral recombination sites while preserving a reading frame with

plasmid based fusions. Primer has several useful features such as sorting

calculated primer sets by target size, facilitating interpretation of PCR

amplifications by agarose gel electrophoresis, as well as supplying the molecular

biologist with many important characteristics of each target such as the expected

size of the PCR amplified DNA fragment and internal restriction sites. The NESG

has cloned over 1500 genes using oligonucleotide primers designed by Primer.

DOI: 10.1023/B:JSFG.0000029238.86387.90

PMID: 15263839 [Indexed for MEDLINE]

3406. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D552-6.

GenePaint.org: an atlas of gene expression patterns in the mouse embryo.

Visel A(1), Thaller C, Eichele G.

Author information:

(1)Max Planck Institute of Experimental Endocrinology, Feodor-Lynen-Strasse 7,

D-30625 Hannover, Germany.

High-throughput instruments were recently developed to determine gene expression

patterns on tissue sections by RNA in situ hybridization. The resulting images of

gene expression patterns, chiefly of E14.5 mouse embryos, are accessible to the

public at http://www.genepaint.org. This relational database is searchable for

gene identifiers and RNA probe sequences. Moreover, patterns and intensity of

expression in approximately 100 different embryonic tissues are annotated and can

be searched using a standardized catalog of anatomical structures. A virtual

microscope tool, the Zoom Image Server, was implemented in GenePaint.org and

permits interactive zooming and panning across approximately 15,000

high-resolution images.

DOI: 10.1093/nar/gkh029

PMCID: PMC308763

PMID: 14681479 [Indexed for MEDLINE]

3407. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D273-6.

FusionDB: a database for in-depth analysis of prokaryotic gene fusion events.

Suhre K(1), Claverie JM.

Author information:

(1)Information Génomique and Structurale, CNRS-UPR 2589, 31 chemin Joseph

Aiguier, 13402 Marseille Cedex 20, France. karsten.suhre@igs.cnrs-mrs.fr

FusionDB (http://igs-server.cnrs-mrs.fr/FusionDB/) constitutes a resource

dedicated to in-depth analysis of bacterial and archaeal gene fusion events. Such

events can provide the 'Rosetta stone' in the search for potential

protein-protein interactions, as well as metabolic and regulatory networks.

However, the false positive rate of this approach may be quite high, prompting a

detailed scrutiny of putative gene fusion events. FusionDB readily provides much

of the information required for that task. Moreover, FusionDB extends the notion

of gene fusion from that of a single gene to that of a family of genes by

assembling pairs of genes from different genomes that belong to the same Cluster

of Orthogonal Groups (COG). Multiple sequence alignments and phylogenetic tree

reconstruction for the N- and C-terminal parts of these 'COG fusion' events are

provided to distinguish single and multiple fusion events from cases of gene

fission, pseudogenes and other false positives. Finally, gene fusion events with

matches to known structures of heterodimers in the Protein Data Bank (PDB) are

identified and may be visualized. FusionDB is fully searchable with access to

sequence and alignment data at all levels. A number of different scores are

provided to easily differentiate 'real' from 'questionable' cases, especially

when larger database searches are performed. FusionDB is cross-linked with the

'Phylogenomic Display of Bacterial Genes' (PhydBac) online web server. Together,

these servers provide the complete set of information required for in-depth

analysis of non-homology-based gene function attribution.

DOI: 10.1093/nar/gkh053

PMCID: PMC308787

PMID: 14681411 [Indexed for MEDLINE]

3408. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D230-4.

The SWISS-MODEL Repository of annotated three-dimensional protein structure

homology models.

Kopp J(1), Schwede T.

Author information:

(1)Biozentrum der Universität Basel and Swiss Institute of Bioinformatics,

Klingelbergstrasse 50-70, CH 4056 Basel, Switzerland.

The SWISS-MODEL Repository is a database of annotated three-dimensional

comparative protein structure models generated by the fully automated

homology-modelling pipeline SWISS-MODEL. The Repository currently contains about

300,000 three-dimensional models for sequences from the Swiss-Prot and TrEMBL

databases. The content of the Repository is updated on a regular basis

incorporating new sequences, taking advantage of new template structures becoming

available and reflecting improvements in the underlying modelling algorithms.

Each entry consists of one or more three-dimensional protein models, the

superposed template structures, the alignments on which the models are based, a

summary of the modelling process and a force field based quality assessment. The

SWISS-MODEL Repository can be queried via an interactive website at

http://swissmodel.expasy. org/repository/. Annotation and cross-linking of the

models with other databases, e.g. Swiss-Prot on the ExPASy server, allow for

seamless navigation between protein sequence and structure information. The aim

of the SWISS-MODEL Repository is to provide access to an up-to-date collection of

annotated three-dimensional protein models generated by automated homology

modelling, bridging the gap between sequence and structure databases.

DOI: 10.1093/nar/gkh008

PMCID: PMC308743

PMID: 14681401 [Indexed for MEDLINE]

3409. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D193-5.

DomIns: a web resource for domain insertions in known protein structures.

Selvam RA(1), Sasidharan R.

Author information:

(1)The Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10

1SA, UK.

Proteins can be formed by single or multiple domains. The process of

recombination at the molecular level has generated a wide variety of multi-domain

proteins with specific domain organization to cater to the functional

requirements of an organism. The functional and structural costs of inserting a

domain into another means that multi-domain proteins are usually formed by

covalently linking the N-terminus of one domain to the C-terminus of the

preceding domain. While this is true in a large proportion of multi-domain

proteins, we find a significant fraction of proteins that are the result of

domain insertion. The inserted domain breaks the sequence contiguity of the

domain into which it is inserted leading to a novel domain organization. This web

resource aims to document domain insertions in known protein structures that are

classified in the SCOP database. The web server can be accessed from

http://stash.mrc-lmb.cam. ac.uk/DomIns/.

DOI: 10.1093/nar/gkh047

PMCID: PMC308781

PMID: 14681392 [Indexed for MEDLINE]

3410. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D104-8.

The tmRNA website: reductive evolution of tmRNA in plastids and other

endosymbionts.

Gueneau de Novoa P(1), Williams KP.

Author information:

(1)Centro de Biofisica y Bioquimica, Instituto Venezolano de Investigaciones

Cientificas, Altos de Pipe, MI, Venezuela.

tmRNA combines tRNA- and mRNA-like properties and ameliorates problems arising

from stalled ribosomes. Research on the mechanism, structure and biology of tmRNA

is served by the tmRNA website (http://www.indiana.edu/~ tmrna), a collection of

sequences, alignments, secondary structures and other information. Because many

of these sequences are not in GenBank, a BLAST server has been added; another new

feature is an abbreviated alignment for the tRNA-like domain only. Many tmRNA

sequences from plastids have been added, five found in public sequence data and

another 10 generated by direct sequencing; detection in early-branching members

of the green plastid lineage brings coverage to all three primary plastid

lineages. The new sequences include the shortest known tmRNA sequence. While

bacterial tmRNAs usually have a lone pseudoknot upstream of the mRNA segment and

a string of three or four pseudoknots downstream, plastid tmRNAs collectively

show loss of pseudoknots at both postions. The pseudoknot-string region is also

too short to contain the usual pseudoknot number in another new entry, the tmRNA

sequence from a bacterial endosymbiont of insect cells, Tremblaya princeps.

Pseudoknots may optimize tmRNA function in free-living bacteria, yet become

dispensible when the endosymbiotic lifestyle relaxes selective pressure for fast

growth.

DOI: 10.1093/nar/gkh102

PMCID: PMC308836

PMID: 14681369 [Indexed for MEDLINE]

3411. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D82-5.

The Eukaryotic Promoter Database EPD: the impact of in silico primer extension.

Schmid CD(1), Praz V, Delorenzi M, Périer R, Bucher P.

Author information:

(1)Swiss Institute of Bioinformatics, Ch. des Boveresses 155, 1066 Epalinges

s/Lausanne, Switzerland.

The Eukaryotic Promoter Database (EPD) is an annotated non-redundant collection

of eukaryotic POL II promoters, experimentally defined by a transcription start

site (TSS). There may be multiple promoter entries for a single gene. The

underlying experimental evidence comes from journal articles and, starting from

release 73, from 5' ESTs of full-length cDNA clones used for so-called in silico

primer extension. Access to promoter sequences is provided by pointers to TSS

positions in nucleotide sequence entries. The annotation part of an EPD entry

includes a description of the type and source of the initiation site mapping

data, links to other biological databases and bibliographic references. EPD is

structured in a way that facilitates dynamic extraction of biologically

meaningful promoter subsets for comparative sequence analysis. Web-based

interfaces have been developed that enable the user to view EPD entries in

different formats, to select and extract promoter sequences according to a

variety of criteria and to navigate to related databases exploiting different

cross-references. Tools for analysing sequence motifs around TSSs defined in EPD

are provided by the signal search analysis server. EPD can be accessed at

http://www.epd. isb-sib.ch.

DOI: 10.1093/nar/gkh122

PMCID: PMC308856

PMID: 14681364 [Indexed for MEDLINE]

3412. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D41-4.

MIPS: analysis and annotation of proteins from whole genomes.

Mewes HW(1), Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, Münsterkötter

M, Pagel P, Strack N, Stümpflen V, Warfsmann J, Ruepp A.

Author information:

(1)Institute for Bioinformatics (MIPS), GSF National Research Center for

Environment and Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany.

w.mewes@gsf.de

The Munich Information Center for Protein Sequences (MIPS-GSF), Neuherberg,

Germany, provides protein sequence-related information based on whole-genome

analysis. The main focus of the work is directed toward the systematic

organization of sequence-related attributes as gathered by a variety of

algorithms, primary information from experimental data together with information

compiled from the scientific literature. MIPS maintains automatically generated

and manually annotated genome-specific databases, develops systematic

classification schemes for the functional annotation of protein sequences and

provides tools for the comprehensive analysis of protein sequences. This report

updates the information on the yeast genome (CYGD), the Neurospora crassa genome

(MNCDB), the database of complete cDNAs (German Human Genome Project, NGFN), the

database of mammalian protein-protein interactions (MPPI), the database of FASTA

homologies (SIMAP), and the interface for the fast retrieval of

protein-associated information (QUIPOS). The Arabidopsis thaliana database, the

rice database, the plant EST databases (MATDB, MOsDB, SPUTNIK), as well as the

databases for the comprehensive set of genomes (PEDANT genomes) are described

elsewhere in the 2003 and 2004 NAR database issues, respectively. All databases

described, and the detailed descriptions of our projects can be accessed through

the MIPS web server (http://mips.gsf.de).

DOI: 10.1093/nar/gkh092

PMCID: PMC308826

PMID: 14681354 [Indexed for MEDLINE]

3413. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D31-4.

DDBJ in the stream of various biological data.

Miyazaki S(1), Sugawara H, Ikeo K, Gojobori T, Tateno Y.

Author information:

(1)Center for Information Biology and DNA Data Bank of Japan, National Institute

of Genetics, Yata, Mishima 411-8540, Japan.

In the past year we at DDBJ (http://www.ddbj.nig. ac.jp) have made a steady

increase in the number of data submissions with a 50.6% increment in the number

of bases or 46.5% increment in the number of entries. Among them the genome data

of man, ascidian and rice hold the top three. Our activity has extended to

providing a tool that enables sequence retrieval using regular expressions, and

to launching our SOAP server and web services to facilitate the acquisition of

proper data and tools from a huge number of biological data resources on websites

worldwide. We have also opened our public gene expression database, CIBEX.

DOI: 10.1093/nar/gkh127

PMCID: PMC308861

PMID: 14681352 [Indexed for MEDLINE]

3414. Pac Symp Biocomput. 2004:387-98.

Protein structure and fold prediction using tree-augmented naive Bayesian

classifier.

Chinnasamy A(1), Sung WK, Mittal A.

Author information:

(1)Department of Computer Science, National University of Singapore, 3 Science

Drive 2, Singapore 117543. arun@ksung.comp.nus.edu.sg

For determining the structure class and fold class of Protein Structure,

computer-based techniques have became essential considering the large volume of

the data. Several techniques based on sequence similarity. Neural Networks, SVMs,

etc have been applied. This paper presents a framework using the Tree-Augmented

Networks (TAN) based on the theory of learning Bayesian networks but with less

restrictive assumptions than the naive Bayesian networks. In order to enhance

TAN's performance, pre-processing of data is done by feature discretization and

post-processing is done by using Mean Probability Voting (MPV) scheme. The

advantage of using Bayesian approach over other learning methods is that the

network structure is intuitive. In addition, one can read off the TAN structure

probabilities to determine the significance of each feature (say, Hydrophobicity)

for each class, which help to further understand the mystery of protein

structure. Experimental results and comparison with other works over two

databases show the effectiveness of our TAN based framework. The idea is

implemented as the BAYESPROT web server and it is available at

http://www-appn.comp.nus.edu.sg/-bioinfo/bayesprot/Default.htm.

PMID: 14992519 [Indexed for MEDLINE]

3415. Stud Health Technol Inform. 2004;109:226-43.

Recommendations of the International Medical Informatics Association (IMIA) on

education in health and medical informatics.

International Medical Informatics Association, Working Group 1: Health and

Medical Informatics Education.

The International Medical Informatics Association (IMIA) agreed on international

recommendations in health informatics / medical informatics education. These

should help to establish courses, course tracks or even complete programs in this

field, to further develop existing educational activities in the various nations

and to support international initiatives concerning education in health and

medical informatics (HMI), particularly international activities in educating HMI

specialists and the sharing of courseware. The IMIA recommendations centre on

educational needs for health care professionals to acquire knowledge and skills

in information processing and information and communication technology. The

educational needs are described as a three-dimensional framework. The dimensions

are: 1) professionals in health care (physicians, nurses, HMI professionals,

...), 2) type of specialisation in health and medical informatics (IT users, HMI

specialists) and 3) stage of career progression (bachelor, master, ...). Learning

outcomes are defined in terms of knowledge and practical skills for health care

professionals in their role (a) as IT user and (b) as HMI specialist.

Recommendations are given for courses/course tracks in HMI as part of educational

programs in medicine, nursing, health care management, dentistry, pharmacy,

public health, health record administration, and informatics/computer science as

well as for dedicated programs in HMI (with bachelor, master or doctor degree).

To support education in HMI, IMIA offers to award a certificate for high quality

HMI education and supports information exchange on programs and courses in HMI

through a WWW server of its Working Group on Health and Medical Informatics

Education (http://www.imia.org/wg1).

PMID: 15718686 [Indexed for MEDLINE]

3416. Stud Health Technol Inform. 2004;105:61-9.

New developments in digital pathology: from telepathology to virtual pathology

laboratory.

Kayser K(1), Kayser G, Radziszowski D, Oehmann A.

Author information:

(1)UICC-TPCC, Institute of Pathology, Charite, Berlin, Germany.

AIMS: To analyse the present status and future development of computerized

diagnostic pathology in terms of work-flow integrative telepathology and virtual

laboratory.

PRESENT STATUS: Telepathology has left its childhood. The technical development

of telepathology is mature, in contrast to that of virtual pathology. Two kinds

of virtual pathology laboratories are emerging: a) those with distributed

pathologists and distributed (>=1) laboratories associated to individual biopsy

stations/surgical theatres, and b) distributed pathologists working in a

centralized laboratory. Both are under technical development. Telepathology can

be used for e-learning and e-training in pathology, as exemplarily demonstrated

on Digital Lung Pathology Pathology (www.pathology-online.org).

FEATURES OF VIRTUAL PATHOLOGY: A virtual pathology institution (mode a) accepts a

complete case with the patient's history, clinical findings, and (pre-selected)

images for first diagnosis. The diagnostic responsibility is that of a

conventional institution. The internet serves as platform for information

transfer, and an open server such as the iPATH (http://telepath.patho.unibas.ch)

for coordination and performance of the diagnostic procedure. The size of images

has to be limited, and usual different magnifications have to be used. A group of

pathologists is "on duty", or selects one member for a predefined duty period.

The diagnostic statement of the pathologist(s) on duty is retransmitted to the

sender with full responsibility. First experiences of a virtual pathology

institution group working with the iPATH server (Dr. L. Banach, Dr. G. Haroske,

Dr. I. Hurwitz, Dr. K. Kayser, Dr. K.D. Kunze, Dr. M. Oberholzer,) working with a

small hospital of the Salomon islands are promising. A centralized virtual

pathology institution (mode b) depends upon the digitalisation of a complete

slide, and the transfer of large sized images to different pathologists working

in one institution. The technical performance of complete slide digitalisation is

still under development and does not completely fulfil the requirements of a

conventional pathology institution at present. VIRTUAL PATHOLOGY AND E-LEARNING:

At present, e-learning systems are "stand-alone" solutions distributed on CD or

via internet. A characteristic example is the Digital Lung Pathology CD

(www.pathology-online.org), which includes about 60 different rare and common

lung diseases and internet access to scientific library systems (PubMed), distant

measurement servers (EuroQuant), or electronic journals (Elec J Pathol Histol). A

new and complete data base based upon this CD will combine e-learning and

e-teaching with the actual workflow in a virtual pathology institution (mode a).

The technological problems are solved and do not depend upon technical

constraints such as slide scanning systems.

PERSPECTIVES: Telepathology serves as promotor for a new landscape in diagnostic

pathology, the so-called virtual pathology institution. Industrial and scientific

efforts will probably allow an implementation of this technique within the next

two years.

PMID: 15718595 [Indexed for MEDLINE]

3417. Technol Health Care. 2004;12(1):33-41.

A first look at HealthCyberMap medical semantic subject search engine.

Boulos MN(1).

Author information:

(1)School for Health, University of Bath, UK. M.N.K.Boulos@bath.ac.uk

HealthCyberMap (http://healthcybermap.semanticweb.org) is a Semantic Web project

that aims at mapping selected parts of health information resources in cyberspace

in novel semantic ways to improve their retrieval and navigation. This paper

describes HealthCyberMap semantic subject search engine methodology and early

prototype which attempt to overcome the limitations of conventional free text

search engines. Explicit concepts in resource metadata map onto a brokering

domain ontology (a clinical terminology or classification) allowing a Semantic

Web search engine to infer implicit meanings (synonyms and semantic

relationships) not directly mentioned in either the resource or its metadata.

Similarly, user queries would map to the same ontology allowing the search engine

to infer the implicit semantics of user queries and use them to optimise

retrieval. Related issues of metadata, clinical terminologies and automatic vs.

manual indexing of medical Web resources are also discussed, together with future

methodological directions, which include the use of a true terminology server as

an intelligent broker between user queries and HealthCyberMap pool of resource

metadata. A comparative evaluation of the new engine based on relevance metrics

is also proposed.

PMID: 15096685 [Indexed for MEDLINE]

3418. Proc Natl Acad Sci U S A. 2003 Dec 23;100(26):15310-5. Epub 2003 Dec 15.

Prediction and statistics of pseudoknots in RNA structures using exactly

clustered stochastic simulations.

Xayaphoummine A(1), Bucher T, Thalmann F, Isambert H.

Author information:

(1)Laboratoire de Dynamique des Fluides Complexes, Centre National de la

Recherche Scientifique-Université Louis Pasteur, Institut de Physique, 3 Rue de

l'Université, 67000 Strasbourg, France.

Ab initio RNA secondary structure predictions have long dismissed helices

interior to loops, so-called pseudoknots, despite their structural importance.

Here we report that many pseudoknots can be predicted through long-time-scale

RNA-folding simulations, which follow the stochastic closing and opening of

individual RNA helices. The numerical efficacy of these stochastic simulations

relies on an O(n2) clustering algorithm that computes time averages over a

continuously updated set of n reference structures. Applying this exact

stochastic clustering approach, we typically obtain a 5- to 100-fold simulation

speed-up for RNA sequences up to 400 bases, while the effective acceleration can

be as high as 105-fold for short, multistable molecules (<or=150 bases). We

performed extensive folding statistics on random and natural RNA sequences and

found that pseudoknots are distributed unevenly among RNA structures and account

for up to 30% of base pairs in G+C-rich RNA sequences (online RNA-folding

kinetics server including pseudoknots: http://kinefold.u-strasbg.fr).

DOI: 10.1073/pnas.2536430100

PMCID: PMC307563

PMID: 14676318 [Indexed for MEDLINE]

3419. BMC Bioinformatics. 2003 Dec 16;4:63.

AnaBench: a Web/CORBA-based workbench for biomolecular sequence analysis.

Badidi E(1), De Sousa C, Lang BF, Burger G.

Author information:

(1)The Canadian Institute for Advanced Research, Program in Evolutionary Biology,

Département de Biochimie, Université de Montréal, 2900, Boul, Edouard Montpetit,

Montréal, QC H3T 1J4, Canada. me.badidi@umontreal.ca

BACKGROUND: Sequence data analyses such as gene identification, structure

modeling or phylogenetic tree inference involve a variety of bioinformatics

software tools. Due to the heterogeneity of bioinformatics tools in usage and

data requirements, scientists spend much effort on technical issues including

data format, storage and management of input and output, and memorization of

numerous parameters and multi-step analysis procedures.

RESULTS: In this paper, we present the design and implementation of AnaBench, an

interactive, Web-based bioinformatics Analysis workBench allowing streamlined

data analysis. Our philosophy was to minimize the technical effort not only for

the scientist who uses this environment to analyze data, but also for the

administrator who manages and maintains the workbench. With new bioinformatics

tools published daily, AnaBench permits easy incorporation of additional tools.

This flexibility is achieved by employing a three-tier distributed architecture

and recent technologies including CORBA middleware, Java, JDBC, and JSP. A CORBA

server permits transparent access to a workbench management database, which

stores information about the users, their data, as well as the description of all

bioinformatics applications that can be launched from the workbench.

CONCLUSION: AnaBench is an efficient and intuitive interactive bioinformatics

environment, which offers scientists application-driven, data-driven and

protocol-driven analysis approaches. The prototype of AnaBench, managed by a team

at the Université de Montréal, is accessible on-line at:

http://malawimonas.bcm.umontreal.ca:8091/anabench. Please contact the authors for

details about setting up a local-network AnaBench site elsewhere.

DOI: 10.1186/1471-2105-4-63

PMCID: PMC328086

PMID: 14678565 [Indexed for MEDLINE]

3420. Bioinformatics. 2003 Dec 12;19(18):2500-1.

ModLoop: automated modeling of loops in protein structures.

Fiser A(1), Sali A.

Author information:

(1)Department of Biochemistry and Seaver Foundation Center for Bioinformatics,

Albert Einstein College of Medicine, 1300 Morris Park Ave., Bronx, NY 10461, USA.

andras@fiserlab.org

SUMMARY: ModLoop is a web server for automated modeling of loops in protein

structures. The input is the atomic coordinates of the protein structure in the

Protein Data Bank format, and the specification of the starting and ending

residues of one or more segments to be modeled, containing no more than 20

residues in total. The output is the coordinates of the non-hydrogen atoms in the

modeled segments. A user provides the input to the server via a simple web

interface, and receives the output by e-mail. The server relies on the loop

modeling routine in MODELLER that predicts the loop conformations by satisfaction

of spatial restraints, without relying on a database of known protein structures.

For a rapid response, ModLoop runs on a cluster of Linux PC computers.

AVAILABILITY: The server is freely accessible to academic users at

http://salilab.org/modloop

PMID: 14668246 [Indexed for MEDLINE]

3421. Bioinformatics. 2003 Dec 12;19(18):2486-8.

PRIMEX: rapid identification of oligonucleotide matches in whole genomes.

Lexa M(1), Valle G.

Author information:

(1)Laboratory of Functional Genomics and Proteomics, Masaryk University Brno,

Kotlarska 2, 61137 Brno, Czech Republic. m-lexa@sci.muni.cz

SUMMARY: PRIMEX (PRImer Match EXtractor) can detect oligonucleotide sequences in

whole genomes, allowing for mismatches. Using a word lookup table and server

functionality, PRIMEX accepts queries from client software and returns matches

rapidly. We find it faster and more sensitive than currently available tools.

AVAILABILITY: Running applications and source code have been made available at

http://bioinformatics.cribi.unipd.it/primex

PMID: 14668240 [Indexed for MEDLINE]

3422. Plant Physiol. 2003 Dec;133(4):1691-701.

Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive

prediction from sequence- and genome-wide studies for Arabidopsis and rice.

Eisenhaber B(1), Wildpaner M, Schultz CJ, Borner GH, Dupree P, Eisenhaber F.

Author information:

(1)Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, A-1030 Vienna,

Austria. Birgit.Eisenhaber@imp.univie.ac.at

Posttranslational glycosylphosphatidylinositol (GPI) lipid anchoring is common

not only for animal and fungal but also for plant proteins. The attachment of the

GPI moiety to the carboxyl-terminus after proteolytic cleavage of a C-terminal

propeptide is performed by the transamidase complex. Its four known subunits also

have obvious full-length orthologs in the Arabidopsis and rice (Oryza sativa)

genomes; thus, the mechanism of substrate protein processing appears similar for

all eukaryotes. A learning set of plant proteins (substrates for the transamidase

complex) has been collected both from the literature and plant sequence

databases. We find that the plant GPI lipid anchor motif differs in minor aspects

from the animal signal (e.g. the plant hydrophobic tail region can contain a

higher fraction of aromatic residues). We have developed the "big-Pi plant"

program for prediction of compatibility of query protein C-termini with the plant

GPI lipid anchor motif requirements. Validation tests show that the sensitivity

for transamidase targets is approximately 94%, and the rate of false positive

prediction is about 0.1%. Thus, the big-Pi predictor can be applied as

unsupervised genome annotation and target selection tool. The program is also

suited for the design of modified protein constructs to test their GPI lipid

anchoring capacity. The big-Pi plant predictor Web server and lists of potential

plant precursor proteins in Swiss-Prot, SPTrEMBL, Arabidopsis, and rice proteomes

are available at http://mendel.imp.univie.ac.at/gpi/plants/gpi\_plants.html.

Arabidopsis and rice protein hits have been functionally classified. Several GPI

lipid-anchored arabinogalactan-related proteins have been identified in rice.

DOI: 10.1104/pp.103.023580

PMCID: PMC300724

PMID: 14681532 [Indexed for MEDLINE]

3423. Bioinformatics. 2003 Nov 22;19(17):2332-3.

MetaGeneAlyse: analysis of integrated transcriptional and metabolite data.

Daub CO(1), Kloska S, Selbig J.

Author information:

(1)Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476

Golm, Germany. daub@mpimp-golm.mpg.de

New techniques in sample preparation allow high throughput analysis of samples on

the transcriptional as well as on the metabolic level. We present a service

accessible via the web that allows the analysis of integrated data sets that

combine gene-expression data and metabolic data. After uploading, data sets can

be normalized, clustered by various methods and results can be graphically

visualized. All calculations are carried out on a server, so even time- and

memory-consuming analyses can be done independently of the performance of the

client.AVAILABILITY: The service is accessible via web-interface at

http://metagenealyse.mpimp-golm.mpg.de/

PMID: 14630670 [Indexed for MEDLINE]

3424. Bioinformatics. 2003 Nov 22;19(17):2294-301.

Efficient remote homology detection using local structure.

Hou Y(1), Hsu W, Lee ML, Bystroff C.

Author information:

(1)School of Computing, National University of Singapore, Singapore 117543.

houyuna@comp.nus.edu.sg

MOTIVATION: The function of an unknown biological sequence can often be

accurately inferred if we are able to map this unknown sequence to its

corresponding homologous family. At present, discriminative methods such as

SVM-Fisher and SVM-pairwise, which combine support vector machine (SVM) and

sequence similarity, are recognized as the most accurate methods, with

SVM-pairwise being the most accurate. However, these methods typically encode

sequence information into their feature vectors and ignore the structure

information. They are also computationally inefficient. Based on these

observations, we present an alternative method for SVM-based protein

classification. Our proposed method, SVM-I-sites, utilizes structure similarity

for remote homology detection.

RESULT: We run experiments on the Structural Classification of Proteins 1.53 data

set. The results show that SVM-I-sites is more efficient than SVM-pairwise.

Further, we find that SVM-I-sites outperforms sequence-based methods such as

PSI-BLAST, SAM, and SVM-Fisher while achieving a comparable performance with

SVM-pairwise.

AVAILABILITY: I-sites server is accessible through the web at

http://www.bioinfo.rpi.edu. Programs are available upon request for academics.

Licensing agreements are available for commercial interests. The framework of

encoding local structure into feature vector is available upon request.

PMID: 14630658 [Indexed for MEDLINE]

3425. BMC Bioinformatics. 2003 Nov 21;4:58.

Online tool for the discrimination of equi-distributions.

Pöschel T(1), Frömmel C, Gille C.

Author information:

(1)Charité, Institut für Biochemie, Monbijoustrasse 2, D-10117, Berlin, Germany.

thorsten.poeschel@charite.de

BACKGROUND: For many applications one wishes to decide whether a certain set of

numbers originates from an equiprobability distribution or whether they are

unequally distributed. Distributions of relative frequencies may deviate

significantly from the corresponding probability distributions due to finite

sample effects. Hence, it is not trivial to discriminate between an

equiprobability distribution and non-equally distributed probabilities when

knowing only frequencies.

RESULTS: Based on analytical results we provide a software tool which allows to

decide whether data correspond to an equiprobability distribution. The tool is

available at http://bioinf.charite.de/equifreq/.

CONCLUSIONS: Its application is demonstrated for the distribution of point

mutations in coding genes.

DOI: 10.1186/1471-2105-4-58

PMCID: PMC317281

PMID: 14629779 [Indexed for MEDLINE]

3426. Comput Methods Programs Biomed. 2003 Nov;72(3):197-208.

A meta-data-based learning resource server for medicine.

Geueke M(1), Stausberg J.

Author information:

(1)Medical Faculty, Institute for Medical Informatics, Biometry and Epidemiology,

University of Essen, Hufelandstr. 55, D-45122 Essen, Germany.

The World Wide Web (WWW) promises many advantages in the distribution and

presentation of electronic learning modules for medical education. However, there

is a well-known obstacle: finding the right offers on demand. We developed a

solution based on international standards called Learning Resource Server

Medicine (LRSMed), which provides a meta-data collection of electronic learning

modules. Based on the meta-data, LRSMed offers a service for the learners to

retrieve WWW resources at the time of learning. We used the IMS Learning Resource

Meta-data Information Model as suitable meta-data specification, the eXtensible

Markup Language (XML) as syntax for interfaces and the Oracle suite for

implementation. LRSMed is available at

http://mmedia.medizin.uni-essen.de/portal/. More than 260 learning resources are

currently registered and described. So far a user-interface has been implemented

which allows searching, commenting and editing. In the next future we will add an

application programming interface (API) for the integration of LRSMed in portals,

learning platforms and information systems of hospitals, medical practices or

healthcare networks.

PMID: 14554134 [Indexed for MEDLINE]

3427. Bioinformatics. 2003 Oct 12;19(15):1985-96.

A computational pipeline for protein structure prediction and analysis at genome

scale.

Shah M(1), Passovets S, Kim D, Ellrott K, Wang L, Vokler I, LoCascio P, Xu D, Xu

Y.

Author information:

(1)Life Sciences Division, Oak Ridge National Laboratory, TN 37830-6480, USA.

MOTIVATION: Experimental techniques alone cannot keep up with the production rate

of protein sequences, while computational techniques for protein structure

predictions have matured to such a level to provide reliable structural

characterization of proteins at large scale. Integration of multiple

computational tools for protein structure prediction can complement experimental

techniques.

RESULTS: We present an automated pipeline for protein structure prediction. The

centerpiece of the pipeline is our threading-based protein structure prediction

system PROSPECT. The pipeline consists of a dozen tools for identification of

protein domains and signal peptide, protein triage to determine the protein type

(membrane or globular), protein fold recognition, generation of atomic structural

models, prediction result validation, etc. Different processing and prediction

branches are determined automatically by a prediction pipeline manager based on

identified characteristics of the protein. The pipeline has been implemented to

run in a heterogeneous computational environment as a client/server system with a

web interface. Genome-scale applications on Caenorhabditis elegans, Pyrococcus

furiosus and three cyanobacterial genomes are presented.

AVAILABILITY: The pipeline is available at

http://compbio.ornl.gov/proteinpipeline/

PMID: 14555633 [Indexed for MEDLINE]

3428. Comput Biol Chem. 2003 Oct;27(4-5):511-9.

The PRALINE online server: optimising progressive multiple alignment on the web.

Simossis VA(1), Heringa J.

Author information:

(1)Bioinformatics Unit, Faculty of Sciences, Vrije Universiteit, De Boelelaan

1081A, 1081 HV Amsterdam, The Netherlands. heringa@cs.vu.nl

We introduce the online server for PRALINE

(http://ibium.cs.vu.nl/programs/pralinewww/), an iterative versatile progressive

multiple sequence alignment (MSA) tool. PRALINE provides various MSA optimisation

strategies including weighted global and local profile pre-processing, secondary

structure-guided alignment and a reliability measure for aligned individual

residue positions. The latter can also be used to optimise the alignment when the

profile pre-processing strategies are iterated. In addition, we have modelled the

server output to enable comprehensive visualisation of the generated alignment

and easy figure generation for publications. The alignment is represented in five

default colour schemes based on: residue type, position conservation, position

reliability, residue hydrophobicity and secondary structure; depending on the

options set. We have also implemented a custom colour scheme that allows the user

to select which colour will represent one or more amino acids in the alignment.

The grouping of sequences, on which the alignment is based, can also be

visualised as a dendrogram. The PRALINE algorithm is designed to work more as a

toolkit for MSA rather than a one step process.

PMID: 14642759 [Indexed for MEDLINE]

3429. Environ Monit Assess. 2003 Oct-Nov;88(1-3):389-97.

Sharing the geo-referenced results of climate change impact research.

Sauchyn DJ(1), Joss BN, Nyirfa WN.

Author information:

(1)Prairie Adaptation Research Collaborative, University of Regina, Regina, SK,

Canada, S4S 7J7. sauchyn@uregina.ca

The Prairie Adaptation Research Collaborative (PARC) has implemented an Internet

Map Server (IMS) at the PARC web site (www.parc.ca) to 1) disseminate the

geo-referenced results of PARC sponsored research on climate change impacts and

adaptation, and 2) address data, information and knowledge management within the

PARC network of researchers and partners. PARC facilitates interdisciplinary

research on adaptation to the impacts of climate change in the Canadian Prairie

Provinces. The web site is intended as a platform for sharing information and

encouraging discussion of climate change impacts and adaptation. The IMS enables

scientists and stakeholders to apply simple climate change scenarios to

geo-referenced biophysical and social data, and dynamically create maps that

display the geographic distribution of potential impacts of climate change. With

a limited capacity for spatial analysis, most geo-processing and the climate

impact modeling is done offline within a GIS environment. The IMS will serve the

output from climate impact models, such that the model results can be customized

by the web site user and be most readily applied to the planning and analysis of

adaptation strategies.

PMID: 14570424 [Indexed for MEDLINE]

3430. J Bioinform Comput Biol. 2003 Oct;1(3):495-504.

BTEVAL: a server for evaluation of beta-turn prediction methods.

Kaur H(1), Raghava GP.

Author information:

(1)Bioinformatics Center, Institute of Microbial Technology, Sector 39A,

Chandigarh, India. raghava@imtech.res.in

This paper describes a web server BTEVAL, developed for assessing the performance

of newly developed beta-turn prediction method and it's ranking with respect to

other existing beta-turn prediction methods. Evaluation of a method can be

carried out on a single protein or a number of proteins. It consists of clean

data set of 426 non-homologous proteins with seven subsets of these proteins.

Users can evaluate their method on any subset or a complete set of data. The

method is assessed at amino acid level and performance is evaluated in terms of

Qtotal, Qpredicted, Qobserved and MCC measures. The server also compares the

performance of the method with other existing beta-turn prediction methods such

as Chou-Fasman algorithm, Thornton's algorithm, GORBTURN, 1-4 and 2-3 Correlation

model, Sequence coupled model and BTPRED. The server is accessible from

http://imtech.res.in/raghava/bteval/

PMID: 15290767 [Indexed for MEDLINE]

3431. J Mol Evol. 2003 Oct;57(4):453-66.

A survey of codon and amino acid frequency bias in microbial genomes focusing on

translational efficiency.

Merkl R(1).

Author information:

(1)Abteilung Molekulare Genetik und Präparative Molekularbiologie, Institut für

Mikrobiologie und Genetik, Göttingen Genomics Laboratory,

Georg-August-Universität Göttingen, Grisebachstrasse 8, D - 37077 G6ttingen,

Germany. rmerkl@gwdg.de

Unequal use of synonymous codons has been found in several prokaryotic and

eukaryotic genomes. This bias has been associated with translational efficiency.

The prevalence of this bias across lineages is currently unknown. Here, a new

method (GCB) to measure codon usage bias is presented. It uses an iterative

approach for the determination of codon scores and allows the computation of an

index of codon bias suitable for interspecies comparison. A server to calculate

GCB-values of individual genes as well as a list of compiled results are

available at www.g21.bio.uni-goettingen.de. The method was applied to complete

bacterial genomes. The relation of codon usage bias with amino acid composition

and the choice of stop codons were determined and discussed.

DOI: 10.1007/s00239-003-2499-1

PMID: 14708578 [Indexed for MEDLINE]

3432. Proteomics. 2003 Oct;3(10):1874-82.

Proteome of Caulobacter crescentus cell cycle publicly accessible on SWICZ

server.

Vohradsky J(1), Janda I, Grünenfelder B, Berndt P, Röder D, Langen H, Weiser J,

Jenal U.

Author information:

(1)Institute of Microbiology, CAS, Prague, Czech Republic. vohr@biomed.cas.cz

Here we present the Swiss-Czech Proteomics Server (SWICZ), which hosts the

proteomic database summarizing information about the cell cycle of the aquatic

bacterium Caulobacter crescentus. The database provides a searchable tool for

easy access of global protein synthesis and protein stability data as examined

during the C. crescentus cell cycle. Protein synthesis data collected from five

different cell cycle stages were determined for each protein spot as a relative

value of the total amount of [(35)S]methionine incorporation. Protein stability

of pulse-labeled extracts were measured during a chase period equivalent to one

cell cycle unit. Quantitative information for individual proteins together with

descriptive data such as protein identities, apparent molecular masses and

isoelectric points, were combined with information on protein function, genomic

context, and the cell cycle stage, and were then assembled in a relational

database with a world wide web interface (http://proteom.biomed.cas.cz), which

allows the database records to be searched and displays the recovered

information. A total of 1250 protein spots were reproducibly detected on

two-dimensional gel electropherograms, 295 of which were identified by mass

spectroscopy. The database is accessible either through clickable two-dimensional

gel electrophoretic maps or by means of a set of dedicated search engines. Basic

characterization of the experimental procedures, data processing, and a

comprehensive description of the web site are presented. In its current state,

the SWICZ proteome database provides a platform for the incorporation of new data

emerging from extended functional studies on the C. crescentus proteome.

DOI: 10.1002/pmic.200300559

PMID: 14625849 [Indexed for MEDLINE]

3433. Sci Justice. 2003 Oct-Dec;43(4):237-48.

An improved protocol for the examination of rogue WWW sites.

Marshall AM(1).

Author information:

(1)Centre for Internet Computing, University of Hull, Scarborough Campus, Filey

Road, Scarborough YO11 3AZ, United Kingdom.

Internet server managers have a range of techniques available to help them

improve service performance and security. These techniques can become barriers to

the investigation of illicit or illegal activity. This paper describes some of

the legitimate techniques which can be used to improve server performance or

security, and which present challenges for the investigator. Furthermore, it

proposes a rigorous procedure which should be followed to ensure that any

investigation of a web site or server has been complete and accurate, and that

all possible useful information has been extracted and examined.

DOI: 10.1016/S1355-0306(03)71783-3

PMID: 14714294

3434. Bioinformatics. 2003 Sep 22;19(14):1837-43.

The design of Jemboss: a graphical user interface to EMBOSS.

Carver T(1), Bleasby A.

Author information:

(1)MRC UK HGMP Resource Centre, Hinxton, Cambridge CB10 1SB, UK.

DESIGN: Jemboss is a graphical user interface (GUI) for the European Molecular

Biology Open Software Suite (EMBOSS). It is being developed at the MRC UK HGMP-RC

as part of the EMBOSS project. This paper explains the technical aspects of the

Jemboss client-server design. The client-server model optionally allows that a

Jemboss user have an account on the remote server. The Jemboss client is written

in Java and is downloaded automatically to a user's workstation via Java Web

Start using the HTML protocol. The client then communicates with the remote

server using SOAP (Simple Object Access Protocol). A Tomcat server listens on the

remote machine and communicates the SOAP requests to a Jemboss server, again

written in Java. This Java server interprets the client requests and executes

them through Java Native Interface (JNI) code written in the C language. Another

C program having setuid privilege, jembossctl, is called by the JNI code to

perform the client requests under the user's account on the server. The commands

include execution of EMBOSS applications, file management and project management

tasks. Jemboss allows the use of JSSE for encryption of communication between the

client and server. The GUI parses the EMBOSS Ajax Command Definition language for

form generation and maximum input flexibility. Jemboss interacts directly with

the EMBOSS libraries to allow dynamic generation of application default settings.

RESULTS: This interface is part of the EMBOSS distribution and has attracted much

interest. It has been set up at many other sites globally as well as being used

at the HGMP-RC for registered users.

AVAILABILITY: The software, EMBOSS and Jemboss, is freely available to academics

and commercial users under the GPL licence. It can be downloaded from the EMBOSS

ftp server: http://www.uk.embnet.org/Software/EMBOSS/,

ftp://ftp.uk.embnet.org/pub/EMBOSS/. Registered HGMP-RC users can access an

installed server from: http://www.uk.embnet.org/Software/EMBOSS/Jemboss/

PMID: 14512356 [Indexed for MEDLINE]

3435. Bioinformatics. 2003 Sep 22;19(14):1748-59.

Recognizing the fold of a protein structure.

Harrison A(1), Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J, Orengo C.

Author information:

(1)Biomolecular Structure and Modelling Unit, Department of Biochemistry and

Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK.

This paper reports a graph-theoretic program, GRATH, that rapidly, and

accurately, matches a novel structure against a library of domain structures to

find the most similar ones. GRATH generates distributions of scores by comparing

the novel domain against the different types of folds that have been classified

previously in the CATH database of structural domains. GRATH uses a measure of

similarity that details the geometric information, number of secondary structures

and number of residues within secondary structures, that any two protein

structures share. Although GRATH builds on well established approaches for

secondary structure comparison, a novel scoring scheme has been introduced to

allow ranking of any matches identified by the algorithm. More importantly, we

have benchmarked the algorithm using a large dataset of 1702 non-redundant

structures from the CATH database which have already been classified into fold

groups, with manual validation. This has facilitated introduction of further

constraints, optimization of parameters and identification of reliable thresholds

for fold identification. Following these benchmarking trials, the correct fold

can be identified with the top score with a frequency of 90%. It is identified

within the ten most likely assignments with a frequency of 98%. GRATH has been

implemented to use via a server

(http://www.biochem.ucl.ac.uk/cgi-bin/cath/Grath.pl). GRATH's speed and accuracy

means that it can be used as a reliable front-end filter for the more accurate,

but computationally expensive, residue based structure comparison algorithm SSAP,

currently used to classify domain structures in the CATH database. With an

increasing number of structures being solved by the structural genomics

initiatives, the GRATH server also provides an essential resource for determining

whether newly determined structures are related to any known structures from

which functional properties may be inferred.

PMID: 14512345 [Indexed for MEDLINE]

3436. Lijec Vjesn. 2003 Sep-Oct;125(9-10):271-4.

[First online continuing tele-education in gastroenterology in Croatia].

[Article in Croatian]

Pulanić R(1), Iveković H, Pulanić D, Vrazić H, Ostojić R, Premuzić M, Lepoglavec

Z.

Author information:

(1)Klinika za unutrasnje bolesti, Medicinski fakultet i KBC Zagreb, Zavod za

gastroenterologiju.

The expansion and popularity of the Internet created the expansion of

tele-medicine, with tele-education as its important part. Such on-line distance

learning is especially important for diseases being in the focus of public health

interest, as diseases of the gastrointestinal, hepatobiliary and pancreatic

system due to their frequency. Therefore, in this study is shown the "TIGEL

project of tele-interventional gastroenterology" that was launched in May 2001 at

the Center for Interventional Gastroenterology, Department of Gastroenterology,

University Department of Medicine, Zagreb University Hospital Center. The project

includes creation of a web site at the server of the Zagreb University School of

Medicine (www.mef.hr/edumed/gastro/index.html), and among the most important

goals of the project is continuous medical tele-education in gastroenterology.

Beside description of the project, one of the founders of continuous on-line

medical education in Croatia, this work describes many advantages but also some

still unsolved questions considering medical tele-education, a very promising but

still developing way of education.

PMID: 15038219 [Indexed for MEDLINE]

3437. Nihon Hoshasen Gijutsu Gakkai Zasshi. 2003 Sep;59(9):1155-63.

[Construction of DICOM-WWW gateway by open source, and application to PDAs using

the high-speed mobile communications network].

[Article in Japanese]

Yokohama N(1).

Author information:

(1)The Wakasa-Wan Energy Research Center, Medical Division.

The author constructed a medical image network system using open source software

that took security into consideration. This system was enabled for search and

browse with a WWW browser, and images were stored in a DICOM server. In order to

realize this function, software was developed to fill in the gap between the

DICOM protocol and HTTP using PHP language. The transmission speed was evaluated

by the difference in protocols between DICOM and HTTP. Furthermore, an attempt

was made to evaluate the convenience of medical image access with a personal

information terminal via the Internet through the high-speed mobile communication

terminal. Results suggested the feasibility of remote diagnosis and application

to emergency care.

PMID: 14593329 [Indexed for MEDLINE]

3438. Pac Health Dialog. 2003 Sep;10(2):178-81.

Pacific Telepathlogy Service at Fiji School of Medicine.

Shashidhar VM(1), Brauchli K, Oberholzer M, Pryor J, Kishore K, Krishna R.

Author information:

(1)Fiji School of Medicine, Suva, Fiji.

Pacific Telepathology service has been established at Fiji school of Medicine

(FSM) with technical support from University of Basel. The service is designed

for remote consultation, continuing medical education (CME) health care research

(HCR). To develop Telepathology services for participation with international

Telepathology community for improving quality of health care in the Pacific.

Telepathology server for "Pacific Pathology Group" has been set up at

http://telepath.patho.unibas.ch/, which is dedicated for Telepathology

consultations bring together health care professionals in the Pacific to overcome

limitations of distance, lack of resources and to improve quality of healthcare

services, Accessed by a computer possessing Internet and email connection,

members send cases and questions, review and comment on other cases and receive

consultation via web or email. Benefits are tremendous in terms of remote

consultation, CME, HCR and improving quality of modern health care even at remote

islands devoid of health care resources. Internet speed or reliability is not a

limiting factor. Virtual institute of pathology has been established in

Switzerland with over 400 Pathologists and is providing consultations to many

countries including Solomon Islands where there is no pathologist. The institute

is functioning efficiently with average reporting turnaround time of 48 hours.

Efficiency is the result of organization and communication. Presently this

Service has been established at the FSM in Fiji Islands primarily for education &

remote consultation with plans to expand to other island countries.

PMID: 18181431 [Indexed for MEDLINE]

3439. Proteins. 2003 Aug 15;52(3):454-65.

Characterization of sequence variability in nucleosome core histone folds.

Sullivan SA(1), Landsman D.

Author information:

(1)Computational Biology Branch, National Center for Biotechnology Information,

National Library of Medicine, National Institutes of Health, Bethesda, Maryland

20894, USA.

The three-helix, approximately 65-residue histone fold domain is the most

structurally conserved part of the core histones H2A, H2B, H3, and H4. However,

it evinces a notable degree of sequence variation within and between histone

classes. We used two approaches to characterize sequence variation in these

histone folds, toward elucidating their structure/function relationships and

evolution. On the one hand we asked how much of the sequence variation seen in

structure-based alignments of the folds maintains physicochemical properties at a

position, and on the other, whether conservation correlates to structural

importance, as measured by the number of residue-to-residue contacts a position

makes. Strong physicochemical conservation or correlation of conservation to

contacts would support the idea that functional constraints, rather than genetic

drift, determines the observed range of variants at a given position. We used an

11-state table of physicochemical properties to classify each position in the

core histone fold (CHF) alignments, and a public website

(http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons\_server.pl)

to score conservation. We found that, depending on histone class, from 38 to 77%

of CHF positions are maximally conserved physicochemically, and that for H2B, H3,

and H4 the degree to which a position is conserved correlates positively to the

number of contacts made by the residue at that position in the crystal structure

of the nucleosome core particle. We also examined the correlation between

conservation and the type of contact (e.g., inter- or intrachain,

histone-histone, or histone-DNA, etc.). For H2B, H3, and H4 we found a positive

correlation between conservation and number of interchain protein contacts. No

such correlation or statistical significance was found for DNA or intrachain

contacts. This suggests that variations in the CHF sequences could be

functionally constrained by requirements to make sufficient interchain histone

contacts. We also suggest that inventory of histone residue variants can augment

functional studies of histones. An example is presented for histone H3.

Copyright 2003 Wiley-Liss, Inc.

DOI: 10.1002/prot.10441

PMID: 12866056 [Indexed for MEDLINE]

3440. Bioinformatics. 2003 Aug 12;19(12):1575-7.

AGenDA: homology-based gene prediction.

Taher L(1), Rinner O, Garg S, Sczyrba A, Brudno M, Batzoglou S, Morgenstern B.

Author information:

(1)International Graduate School for Bioinformatics and Genome Research

University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany.

ltaher@TechFak.Uni-Bielefeld.DE

We present a www server for homology-based gene prediction. The user enters a

pair of evolutionary related genomic sequences, for example from human and mouse.

Our software system uses CHAOS and DIALIGN to calculate an alignment of the input

sequences and then searches for conserved splicing signals and start/stop codons

around regions of local sequence similarity. This way, candidate exons are

identified that are used, in turn, to calculate optimal gene models. The server

returns the constructed gene model by email, together with a graphical

representation of the underlying genomic alignment.

PMID: 12912840 [Indexed for MEDLINE]

3441. DNA Seq. 2003 Aug;14(4):327-30.

The DNA sequence quality machine at IFOM: a simple Web-based tool for

quantitative assessment of sequencing reactions.

Lassandro L(1), Lyons R, Volorio S, Bernard L, Guffanti A.

Author information:

(1)University of Michigan, USA.

DNA sequence quality is a factor of paramount importance in the world of modern

genetic and genomics. Both the sequencing of Human Genome in the "post-draft" era

[NHGRI Standard for quality of Human Genomic Sequences, Rev. 7 July (2002) where

http://www.nhgri.nih.gov/Grant\_info/Funding/ Statements/RFA/quality\_standard.html

is the HTTP address] and recent "high-throughput" approaches to genetic

investigation such as SAGE [Velculescu, V.E., Zhang, L., Vogelstein, B. et al.

(1995) "Serial analysis of gene expression", Science 270, 484-487] need a

reliable, standardized measure of the quality of a sequencing reaction. The

increasing importance of SNP studies also requires a stronger quality control on

sequencing reactions by the final user. We propose here a simple, web-based tool

for integrated sequence quality evaluation, high quality region quantitative

value calculation and chromatogram display. This software is aimed at the small

to medium DNA sequence laboratory or to the single researcher, interested in

getting a quantitative measure of the sequence quality, browsing the chromatogram

and checking the quality values base by base. The program is freely available

from the IFOM bioinformatics web Server at

http://bio.ifom-firc.it/Phred20/index.html.

PMID: 14631655 [Indexed for MEDLINE]

3442. Hybrid Hybridomics. 2003 Aug;22(4):229-34.

Prediction of promiscuous and high-affinity mutated MHC binders.

Bhasin M(1), Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39A,

Chandigarh, India.

The identification of peptides in an antigenic sequence that can bind with high

affinity to a wide range of MHC alleles is one of the challenges in subunit

vaccine design. The mutation of natural peptides is an alternative to obtaining

peptides that can bind to a wide range of MHC alleles with high affinity. A large

number of experiments are typically necessary to identify mutations that define

high-affinity binding peptides. Therefore there is a need to develop a

computational method for detecting amino acid mutations in a peptide for making

it high-affinity or promiscuous MHC binders. This report describes a

high-throughput computer driven solution for the identification of promiscuous

and high-affinity mutated binders of 47 MHC class I alleles by introducing

mutations in an antigenic sequence. The method implements quantitative matrices

for creating optimal mutations in an antigenic sequence. It has two major

options: (i) prediction of promiscuous MHC binders and (ii) prediction of

high-affinity binders. In case of prediction of promiscuous binders, the server

allows a user to select (i) permissible mutations in a peptide; (ii) MHC alleles

to whom it should bind; and (iii) positions at which mutation is allowed. In the

case of prediction of high-affinity binders, the server allows users to specify

the positions that should be conserved in the native protein. In both cases, the

method computes the type of mutations and position of mutations in 9-mer peptides

required to have the desired results. The web server MMBPred is available at

www.imtech.res.in/raghava/mmbpred/.

DOI: 10.1089/153685903322328956

PMID: 14511568 [Indexed for MEDLINE]

3443. Int J Hyg Environ Health. 2003 Aug;206(4-5):437-45.

Toxicity characterization of environmental chemicals by the US National

Toxicology Program: an overview.

Chhabra RS(1), Bucher JR, Wolfe M, Portier C.

Author information:

(1)National Institute of Environmental Health Sciences, Research Triangle Park,

NC 27709, USA. chhabraR@niehs.nih.gov

The US National Toxicology Program (NTP) is an interagency program whose mission

is to evaluate agents of public health concern by developing and applying the

tools of modern toxicology and molecular biology. Chemicals substances or

physical agents selected for toxicology and carcinogenesis evaluations by the NTP

are usually studied in a series of subacute (14-day exposure), subchronic (90-day

exposure) and chronic (2-year exposure) studies in rodents. The NTP has published

more than 500 reports of the findings and conclusions from its toxicology and

carcinogenesis studies. In more specialized studies, the NTP also evaluates

adverse effects on the structure and function of the immune, reproductive,

nervous, and respiratory systems. The program attempts to evaluate and

appropriately incorporate new technologies to improve the way we study the

toxicity of chemicals. For example, the program has extensively evaluated several

transgenic mouse models for their potential use as short-term cancer screens and

has been a full participant in an international effort to examine their

usefulness in pharmaceutical registration. Toxicogenomics, an emerging scientific

field that examines the expression of thousands of genes simultaneously in

response to chemical exposure, holds promise for future application to better

understand the underlying mechanisms of chemical toxicity. A number of public

health issues being addressed by the NTP are not only of national importance but

also have global impact, such as the potential for endocrine disruptors to

influence development and carcinogenesis and the safety of herbal medicines and

dietary supplements. The program participates in the preparation of national and

international toxicity testing guidelines and the findings from NTP studies are

widely used for risk assessments by international organizations and federal

agencies. The NTP maintains databases that contain toxicity, and health and

safety information on a large number of chemicals. These databases are available

from the NTP web site (http://ntp-server.niehs.nih.gov) and are accessed over

100000 times a month from around the world.

DOI: 10.1078/1438-4639-00240

PMID: 12971699 [Indexed for MEDLINE]

3444. J Biomol Struct Dyn. 2003 Aug;21(1):99-109.

Gene recognition from questionable ORFs in bacterial and archaeal genomes.

Chen LL(1), Zhang CT.

Author information:

(1)Department of Physics, Tianjin University, Tianjin, 300072, China.

The ORFs of microbial genomes in annotation files are usually classified into two

groups: the first corresponds to known genes; whereas the second includes

'putative', 'probable', 'conserved hypothetical', 'hypothetical', 'unknown' and

'predicted' ORFs etc. Since the annotation is not 100% accurate, it is essential

to confirm which ORF of the latter group is coding and which is not. Starting

from known genes in the former, this paper describes an improved Z curve method

to recognize genes in the latter. Ten-fold cross-validation tests show that the

average accuracy of the algorithm is greater than 99% for recognizing the known

genes in 57 bacterial and archaeal genomes. The method is then applied to

recognize genes of the latter group. The likely non-coding ORFs in each of the 57

bacterial or archaeal genomes studied here are recognized and listed at the

website http://tubic.tju.edu.cn/ZCURVE\_C\_html/noncoding.html. The working

mechanism of the algorithm has been discussed in details. A computer program,

called ZCURVE\_C, was written to calculate a coding score called Z-curve score for

ORFs in the above 57 bacterial and archaeal genomes. Coding/non-coding is simply

determined by the criterion of Z-curve score > 0/ Z-curve score < 0. A website

has been set up to provide the service to calculate the Z-curve score. A user may

submit the DNA sequence of an ORF to the server at

http://tubic.tju.edu.cn/ZCURVE\_C/Default.cgi, and the Z-curve score of the ORF is

calculated and returned to the user immediately.

DOI: 10.1080/07391102.2003.10506908

PMID: 12854962 [Indexed for MEDLINE]

3445. Protein Sci. 2003 Aug;12(8):1652-62.

Prediction of lipoprotein signal peptides in Gram-negative bacteria.

Juncker AS(1), Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A.

Author information:

(1)Center for Biological Sequence Analysis, Technical University of Denmark,

Lyngby 2800, Denmark.

A method to predict lipoprotein signal peptides in Gram-negative Eubacteria,

LipoP, has been developed. The hidden Markov model (HMM) was able to distinguish

between lipoproteins (SPaseII-cleaved proteins), SPaseI-cleaved proteins,

cytoplasmic proteins, and transmembrane proteins. This predictor was able to

predict 96.8% of the lipoproteins correctly with only 0.3% false positives in a

set of SPaseI-cleaved, cytoplasmic, and transmembrane proteins. The results

obtained were significantly better than those of previously developed methods.

Even though Gram-positive lipoprotein signal peptides differ from Gram-negatives,

the HMM was able to identify 92.9% of the lipoproteins included in a

Gram-positive test set. A genome search was carried out for 12 Gram-negative

genomes and one Gram-positive genome. The results for Escherichia coli K12 were

compared with new experimental data, and the predictions by the HMM agree well

with the experimentally verified lipoproteins. A neural network-based predictor

was developed for comparison, and it gave very similar results. LipoP is

available as a Web server at www.cbs.dtu.dk/services/LipoP/.

DOI: 10.1110/ps.0303703

PMCID: PMC2323952

PMID: 12876315 [Indexed for MEDLINE]

3446. Bioinformatics. 2003 Jul 22;19(11):1381-90.

Identifying property based sequence motifs in protein families and superfamilies:

application to DNase-1 related endonucleases.

Mathura VS(1), Schein CH, Braun W.

Author information:

(1)Sealy Center for Structural Biology, Department of Human Biological Chemistry

and Genetics, University of Texas Medical Branch, Galveston, TX 77555-1157, USA.

MOTIVATION: Identification of short conserved sequence motifs common to a protein

family or superfamily can be more useful than overall sequence similarity in

suggesting the function of novel gene products. Locating motifs still requires

expert knowledge, as automated methods using stringent criteria may not

differentiate subtle similarities from statistical noise.

RESULTS: We have developed a novel automatic method, based on patterns of

conservation of 237 physical-chemical properties of amino acids in aligned

protein sequences, to find related motifs in proteins with little or no overall

sequence similarity. As an application, our web-server MASIA identified 12

property-based motifs in the apurinic/apyrimidinic endonuclease (APE) family of

DNA-repair enzymes of the DNase-I superfamily. Searching with these motifs

located distantly related representatives of the DNase-I superfamily, such as

Inositol 5'-polyphosphate phosphatases in the ASTRAL40 database, using a Bayesian

scoring function. Other proteins containing APE motifs had no overall sequence or

structural similarity. However, all were phosphatases and/or had a metal ion

binding active site. Thus our automated method can identify discrete elements in

distantly related proteins that define local structure and aspects of function.

We anticipate that our method will complement existing ones to functionally

annotate novel protein sequences from genomic projects.

AVAILABILITY: MASIA WEB site: http://www.scsb.utmb.edu/masia/masia.html

SUPPLEMENTARY INFORMATION: The dendrogram of 42 APE sequences used to derive

motifs is available on

http://www.scsb.utmb.edu/comp\_biol.html/DNA\_repair/publication.html

PMID: 12874050 [Indexed for MEDLINE]

3447. Bioinformatics. 2003 Jul 1;19(10):1290-1.

The DynDom database of protein domain motions.

Lee RA(1), Razaz M, Hayward S.

Author information:

(1)Royal Society-Wolfson Bioinformatics Laboratory, School of Computing Sciences,

University of East Anglia, Norwich NR4 7TJ, UK.

A relational database has been developed based on the results from the

application of the DynDom program to a number of proteins for which multiple

X-ray conformers are available. The database is populated via a web-based tool

that allows visitors to the website to run the DynDom program server-side by

selecting pairs of X-ray conformers by Protein Data Bank code and chain

identifier.AVAILABILITY: The website can be found at:

http://www.sys.uea.ac.uk/dyndom.

PMID: 12835274 [Indexed for MEDLINE]

3448. J Biomol NMR. 2003 Jul;26(3):215-40.

Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts.

Neal S(1), Nip AM, Zhang H, Wishart DS.

Author information:

(1)Faculty of Pharmacy & Pharmaceutical Sciences, University of Alberta,

Edmonton, AB T6G 2N8, Canada.

A computer program (SHIFTX) is described which rapidly and accurately calculates

the diamagnetic 1H, 13C and 15N chemical shifts of both backbone and sidechain

atoms in proteins. The program uses a hybrid predictive approach that employs

pre-calculated, empirically derived chemical shift hypersurfaces in combination

with classical or semi-classical equations (for ring current, electric field,

hydrogen bond and solvent effects) to calculate 1H, 13C and 15N chemical shifts

from atomic coordinates. The chemical shift hypersurfaces capture dihedral angle,

sidechain orientation, secondary structure and nearest neighbor effects that

cannot easily be translated to analytical formulae or predicted via classical

means. The chemical shift hypersurfaces were generated using a database of

IUPAC-referenced protein chemical shifts--RefDB (Zhang et al., 2003), and a

corresponding set of high resolution (<2.1 A) X-ray structures. Data mining

techniques were used to extract the largest pairwise contributors (from a list of

approximately 20 derived geometric, sequential and structural parameters) to

generate the necessary hypersurfaces. SHIFTX is rapid (<1 CPU second for a

complete shift calculation of 100 residues) and accurate. Overall, the program

was able to attain a correlation coefficient (r) between observed and calculated

shifts of 0.911 (1Halpha), 0.980 (13Calpha), 0.996 (13Cbeta), 0.863 (13CO), 0.909

(15N), 0.741 (1HN), and 0.907 (sidechain 1H) with RMS errors of 0.23, 0.98, 1.10,

1.16, 2.43, 0.49, and 0.30 ppm, respectively on test data sets. We further show

that the agreement between observed and SHIFTX calculated chemical shifts can be

an extremely sensitive measure of the quality of protein structures. Our results

suggest that if NMR-derived structures could be refined using heteronuclear

chemical shifts calculated by SHIFTX, their precision could approach that of the

highest resolution X-ray structures. SHIFTX is freely available as a web server

at http://redpoll.pharmacy.ualberta.ca.

PMID: 12766419 [Indexed for MEDLINE]

3449. Nucleic Acids Res. 2003 Jul 1;31(13):3866-8.

Update on XplorMed: A web server for exploring scientific literature.

Perez-Iratxeta C(1), Pérez AJ, Bork P, Andrade MA.

Author information:

(1)European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg,

Germany. cperez@embl-heidelberg.de

As scientific literature databases like MEDLINE increase in size, so does the

time required to search them. Scientists must frequently inspect long lists of

references manually, often just reading the titles. XplorMed is a web tool that

aids MEDLINE searching by summarizing the subjects contained in the results, thus

allowing users to focus on subjects of interest. Here we describe new features

added to XplorMed during the last 2 years

(http://www.bork.embl-heidelberg.de/xplormed/).

PMCID: PMC168945

PMID: 12824439 [Indexed for MEDLINE]

3450. Nucleic Acids Res. 2003 Jul 1;31(13):3836-9.

Biological SOAP servers and web services provided by the public sequence data

bank.

Sugawara H(1), Miyazaki S.

Author information:

(1)Center for Information Biology and DNA Data Bank of Japan, National Institute

of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan.

hsugawar@genes.nig.uc.jp

A number of biological data resources (i.e. databases and data analytical tools)

are searchable and usable on-line thanks to the internet and the World Wide Web

(WWW) servers. The output from the web server is easy for us to browse. However,

it is laborious and sometimes impossible for us to write a computer program that

finds a useful data resource, sends a proper query and processes the output. It

is a serious obstacle to the integration of distributed heterogeneous data

resources. To solve the issue, we have implemented a SOAP (Simple Object Access

Protocol) server and web services that provide a program-friendly interface. The

web services are accessible at http://www.xml.nig.ac.jp/.

PMCID: PMC168965

PMID: 12824432 [Indexed for MEDLINE]

3451. Nucleic Acids Res. 2003 Jul 1;31(13):3829-32.

PipeAlign: A new toolkit for protein family analysis.

Plewniak F(1), Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel

T, Moulinier L, Muller A, Muller J, Prigent V, Ripp R, Thierry JC, Thompson JD,

Wicker N, Poch O.

Author information:

(1)Laboratoire de Biologie et Génomique Structurales, Institut de Génétique et de

Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP 10142, 67404 Illkirch

Cedex, France.

PipeAlign is a protein family analysis tool integrating a five step process

ranging from the search for sequence homologues in protein and 3D structure

databases to the definition of the hierarchical relationships within and between

subfamilies. The complete, automatic pipeline takes a single sequence or a set of

sequences as input and constructs a high-quality, validated MACS (multiple

alignment of complete sequences) in which sequences are clustered into potential

functional subgroups. For the more experienced user, the PipeAlign server also

provides numerous options to run only a part of the analysis, with the

possibility to modify the default parameters of each software module. For

example, the user can choose to enter an existing multiple sequence alignment for

refinement, validation and subsequent clustering of the sequences. The aim is to

provide an interactive workbench for the validation, integration and presentation

of a protein family, not only at the sequence level, but also at the structural

and functional levels. PipeAlign is available at

http://igbmc.u-strasbg.fr/PipeAlign/.

PMCID: PMC168925

PMID: 12824430 [Indexed for MEDLINE]

3452. Nucleic Acids Res. 2003 Jul 1;31(13):3824-8.

Comprehensive quantitative analyses of the effects of promoter sequence elements

on mRNA transcription.

Lapidot M(1), Pilpel Y.

Author information:

(1)Department of Molecular Genetics, Weizmann Institute of Science, Rehovot,

76100, Israel.

We have generated a WWW interface for automated comprehensive analyses of

promoter regulatory motifs and the effect they exert on mRNA expression profiles.

The server provides a wide spectrum of analysis tools that allow de novo

discovery of regulatory motifs, along with refinement and in-depth investigation

of fully or partially characterized motifs. The presented discovery and analysis

tools are fundamentally different from existing tools in their basic rational,

statistical background and specificity and sensitivity towards true regulatory

elements. We thus anticipate that the service will be of great importance to the

experimental and computational biology communities alike. The motif discovery and

diagnosis workbench is available at http://longitude.weizmann.ac.il/rMotif/.

PMCID: PMC168999

PMID: 12824429 [Indexed for MEDLINE]

3453. Nucleic Acids Res. 2003 Jul 1;31(13):3804-7.

ORFeus: Detection of distant homology using sequence profiles and predicted

secondary structure.

Ginalski K(1), Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L.

Author information:

(1)Bioinformatics Laboratory, BioInfoBank Institute, ul. Limanowskiego 24A,

60-744 Poznan, Poland.

ORFeus is a fully automated, sensitive protein sequence similarity search server

available to the academic community via the Structure Prediction Meta Server

(http://BioInfo.PL/Meta/). The goal of the development of ORFeus was to increase

the sensitivity of the detection of distantly related protein families. Predicted

secondary structure information was added to the information about sequence

conservation and variability, a technique known from hybrid threading approaches.

The accuracy of the meta profiles created this way is compared with profiles

containing only sequence information and with the standard approach of aligning a

single sequence with a profile. Additionally, the alignment of meta profiles is

more sensitive in detecting remote homology between protein families than if

aligning two sequence-only profiles or if aligning a profile with a sequence. The

specificity of the alignment score is improved in the lower specificity range

compared with the robust sequence-only profiles.

PMCID: PMC168911

PMID: 12824423 [Indexed for MEDLINE]

3454. Nucleic Acids Res. 2003 Jul 1;31(13):3795-8.

WU-Blast2 server at the European Bioinformatics Institute.

Lopez R(1), Silventoinen V, Robinson S, Kibria A, Gish W.

Author information:

(1)European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton,

Cambridge, CB10 1SD, UK. rls@ebi.ac.uk

Since 1995, the WU-BLAST programs (http://blast.wustl.edu) have provided a fast,

flexible and reliable method for similarity searching of biological sequence

databases. The software is in use at many locales and web sites. The European

Bioinformatics Institute's WU-Blast2 (http://www.ebi.ac.uk/blast2/) server has

been providing free access to these search services since 1997 and today supports

many features that both enhance the usability and expand on the scope of the

software.

PMCID: PMC168979

PMID: 12824421 [Indexed for MEDLINE]

3455. Nucleic Acids Res. 2003 Jul 1;31(13):3792-4.

BLAST2SRS, a web server for flexible retrieval of related protein sequences in

the SWISS-PROT and SPTrEMBL databases.

Bimpikis K(1), Budd A, Linding R, Gibson TJ.

Author information:

(1)European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg,

Germany.

SRS (Sequence Retrieval System) is a widely used keyword search engine for

querying biological databases. BLAST2 is the most widely used tool to query

databases by sequence similarity search. These tools allow users to retrieve

sequences by shared keyword or by shared similarity, with many public web servers

available. However, with the increasingly large datasets available it is now

quite common that a user is interested in some subset of homologous sequences but

has no efficient way to restrict retrieval to that set. By allowing the user to

control SRS from the BLAST output, BLAST2SRS (http://blast2srs.embl.de/) aims to

meet this need. This server therefore combines the two ways to search sequence

databases: similarity and keyword.

PMCID: PMC168942

PMID: 12824420 [Indexed for MEDLINE]

3456. Nucleic Acids Res. 2003 Jul 1;31(13):3784-8.

ExPASy: The proteomics server for in-depth protein knowledge and analysis.

Gasteiger E(1), Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A.

Author information:

(1)Swiss Institute of Bioinformatics, Centre Médical Universitaire, 1 Rue Michel

Servet, 1211 Geneva 4, Switzerland. elisabeth.gasteiger@isb-sib.ch

The ExPASy (the Expert Protein Analysis System) World Wide Web server

(http://www.expasy.org), is provided as a service to the life science community

by a multidisciplinary team at the Swiss Institute of Bioinformatics (SIB). It

provides access to a variety of databases and analytical tools dedicated to

proteins and proteomics. ExPASy databases include SWISS-PROT and TrEMBL,

SWISS-2DPAGE, PROSITE, ENZYME and the SWISS-MODEL repository. Analysis tools are

available for specific tasks relevant to proteomics, similarity searches, pattern

and profile searches, post-translational modification prediction, topology

prediction, primary, secondary and tertiary structure analysis and sequence

alignment. These databases and tools are tightly interlinked: a special emphasis

is placed on integration of database entries with related resources developed at

the SIB and elsewhere, and the proteomics tools have been designed to read the

annotations in SWISS-PROT in order to enhance their predictions. ExPASy started

to operate in 1993, as the first WWW server in the field of life sciences. In

addition to the main site in Switzerland, seven mirror sites in different

continents currently serve the user community.

PMCID: PMC168970

PMID: 12824418 [Indexed for MEDLINE]

3457. Nucleic Acids Res. 2003 Jul 1;31(13):3782-3.

Swiss EMBnet node web server.

Falquet L(1), Bordoli L, Ioannidis V, Pagni M, Jongeneel CV.

Author information:

(1)Swiss Institute of Bioinformatics, CH-1066 Epalinges, Lausanne, Switzerland.

laurent.falquet@isb-sib.ch

EMBnet is a consortium of collaborating bioinformatics groups located mainly

within Europe (http://www.embnet.org). Each member country is represented by a

'node', a group responsible for the maintenance of local services for their users

(e.g. education, training, software, database distribution, technical support,

helpdesk). Among these services a web portal with links and access to locally

developed and maintained software is essential and different for each node. Our

web portal targets biomedical scientists in Switzerland and elsewhere, offering

them access to a collection of important sequence analysis tools mirrored from

other sites or developed locally. We describe here the Swiss EMBnet node web site

(http://www.ch.embnet.org), which presents a number of original services not

available anywhere else.

PMCID: PMC168954

PMID: 12824417 [Indexed for MEDLINE]

3458. Nucleic Acids Res. 2003 Jul 1;31(13):3771-4.

SIRW: A web server for the Simple Indexing and Retrieval System that combines

sequence motif searches with keyword searches.

Ramu C(1).

Author information:

(1)European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg,

Germany. chenna@embl.de

SIRW (http://sirw.embl.de/) is a World Wide Web interface to the Simple Indexing

and Retrieval System (SIR) that is capable of parsing and indexing various flat

file databases. In addition it provides a framework for doing sequence analysis

(e.g. motif pattern searches) for selected biological sequences through keyword

search. SIRW is an ideal tool for the bioinformatics community for searching as

well as analyzing biological sequences of interest.

PMCID: PMC168953

PMID: 12824415 [Indexed for MEDLINE]

3459. Nucleic Acids Res. 2003 Jul 1;31(13):3767-70.

RNA-related tools on the Bielefeld Bioinformatics Server.

Sczyrba A(1), Krüger J, Mersch H, Kurtz S, Giegerich R.

Author information:

(1)Technische Fakultät, Universität Bielefeld, Postfach 10 01 31, D-33501

Bielefeld. Zentrum für Bioinformatik, Universität Hamburg, Bundesstrasse 43,

D-20146 Hamburg, Germany.

We present four tools for the analysis of RNA secondary structure. They provide

animated visualization of multiple structures, prediction of potential

conformational switching, structure comparison (including local structure

alignment) and prediction of structures potentially containing a certain kind of

pseudoknots. All are available via the Bielefeld University Bioinformatics Server

(http://bibiserv.techfak.uni-bielefeld.de).

PMCID: PMC168982

PMID: 12824414 [Indexed for MEDLINE]

3460. Nucleic Acids Res. 2003 Jul 1;31(13):3736-7.

Bioverse: Functional, structural and contextual annotation of proteins and

proteomes.

McDermott J(1), Samudrala R.

Author information:

(1)Computational Genomics Group, Department of Microbiology, University of

Washington School of Medicine, Seattle, WA 98195, USA.

Functional annotation is routinely performed for large-scale genomics projects

and databases. Researchers working on more specific problems, for instance on an

individual pathway or complex, also need to be able to quickly, completely and

accurately annotate sequences. The Bioverse sequence annotation server

(http://bioverse.compbio.washington.edu) provides a web-based interface to allow

users to submit protein sequences to the Bioverse framework. Sequences are

functionally and structurally annotated and potential contextual annotations are

provided. Researchers can also submit candidate genomes for annotation of all

proteins encoded by the genome (proteome).

PMCID: PMC168957

PMID: 12824406 [Indexed for MEDLINE]

3461. Nucleic Acids Res. 2003 Jul 1;31(13):3720-2.

Phydbac (phylogenomic display of bacterial genes): An interactive resource for

the annotation of bacterial genomes.

Enault F(1), Suhre K, Poirot O, Abergel C, Claverie JM.

Author information:

(1)Structural and Genomic Information (CNRS UPR2589), Institut de Biologie

Structurale et Microbiologie, 31 chemin Joseph Aiguier, 13402 Marseille, cedex

20, France. enault@igs.cnrs-mrs.fr

Phydbac is a web interactive resource based on phylogenomic profiling, designed

to help microbiologists to annotate bacterial proteins. Phylogenomic annotation

is based on the assumption that functionally linked protein-coding genes must

evolve in a coordinated manner. The detection of subsets of co-evolving genes

within a given genome involves the computation of protein sequence conservation

profiles across a spectrum of microbial species, followed by the identification

of significant pairwise correlations between them. Many ongoing studies are

devoted to the problem of computing the most biologically significant

phylogenomic profiles and how best identifying clusters of 'functionally

interacting' genes. Here we introduce a web tool, Phydbac, allowing the dynamic

construction of phylogenomic profiles of protein sequences of interest and their

interactive display. In addition, Phydbac can identify Escherichia coli proteins

exhibiting the evolution pattern most similar to arbitrary query protein

sequences, hence providing functional hints for open reading frames (ORFs) of

hypothetical or unknown function. The phylogenomic profiles of all E.coli K-12

protein-coding genes are pre-computed, allowing queries about E.coli genes to be

answered instantaneously. The profiles and phylogenomic neighborhoods are

computed using an original method shown to perform better than previous ones. An

extension of Phydbac, including precomputed profiles for all available bacterial

genomes (including major pathogens) will soon be available. Phydbac can be

accessed at: http://igs-server.cnrs-mrs.fr/phydbac/.

PMCID: PMC169009

PMID: 12824402 [Indexed for MEDLINE]

3462. Nucleic Acids Res. 2003 Jul 1;31(13):3712-5.

Automated Gene Ontology annotation for anonymous sequence data.

Hennig S(1), Groth D, Lehrach H.

Author information:

(1)Max-Planck Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin,

Germany. hennig@molgen.mpg.de

Gene Ontology (GO) is the most widely accepted attempt to construct a unified and

structured vocabulary for the description of genes and their products in any

organism. Annotation by GO terms is performed in most of the current genome

projects, which besides generality has the advantage of being very convenient for

computer based classification methods. However, direct use of GO in small

sequencing projects is not easy, especially for species not commonly represented

in public databases. We present a software package (GOblet), which performs

annotation based on GO terms for anonymous cDNA or protein sequences. It uses the

species independent GO structure and vocabulary together with a series of protein

databases collected from various sites, to perform a detailed GO annotation by

sequence similarity searches. The sensitivity and the reference protein sets can

be selected by the user. GOblet runs automatically and is available as a public

service on our web server. The paper also addresses the reliability of automated

GO annotations by using a reference set of more than 6000 human proteins. The

GOblet server is accessible at http://goblet.molgen.mpg.de.

PMCID: PMC168988

PMID: 12824400 [Indexed for MEDLINE]

3463. Nucleic Acids Res. 2003 Jul 1;31(13):3698-700.

BPROMPT: A consensus server for membrane protein prediction.

Taylor PD(1), Attwood TK, Flower DR.

Author information:

(1)Edward Jenner Institute for Vaccine Research, Bioinformatics, Compton,

Berkshire, UK. paul.taylor@jenner.ac.uk

Protein structure prediction is a cornerstone of bioinformatics research.

Membrane proteins require their own prediction methods due to their intrinsically

different composition. A variety of tools exist for topology prediction of

membrane proteins, many of them available on the Internet. The server described

in this paper, BPROMPT (Bayesian PRediction Of Membrane Protein Topology), uses a

Bayesian Belief Network to combine the results of other prediction methods,

providing a more accurate consensus prediction. Topology predictions with

accuracies of 70% for prokaryotes and 53% for eukaryotes were achieved. BPROMPT

can be accessed at http://www.jenner.ac.uk/BPROMPT.

PMCID: PMC168961

PMID: 12824397 [Indexed for MEDLINE]

3464. Nucleic Acids Res. 2003 Jul 1;31(13):3688-91.

NEBcutter: A program to cleave DNA with restriction enzymes.

Vincze T(1), Posfai J, Roberts RJ.

Author information:

(1)New England Biolabs, Inc., 32 Tozer Road, Beverly, MA 01915, USA.

NEBcutter, version 1.0, is a program available via a web server

(http://tools.neb.com/NEBcutter) that will accept an input DNA sequence and

produce a comprehensive report of the restriction enzymes that will cleave the

sequence. It produces a variety of outputs including restriction enzyme maps,

theoretical digests and links into the restriction enzyme database, REBASE

(http://www.neb.com/rebase). Importantly, its table of recognition sites is

updated daily from REBASE and it marks all sites that are potentially affected by

DNA methylation (Dam, Dcm, etc.). Many options exist to choose the enzymes used

for digestion, including all known specificities, subsets of those that are

commercially available or sets of enzymes that produce compatible termini.

PMCID: PMC168933

PMID: 12824395 [Indexed for MEDLINE]

3465. Nucleic Acids Res. 2003 Jul 1;31(13):3686-7.

DNA analysis servers: plot.it, bend.it, model.it and IS.

Vlahovicek K(1), Kaján L, Pongor S.

Author information:

(1)Protein Structure and Bioinformatics Group, International Centre for Genetic

Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy.

The WWW servers at http://www.icgeb.trieste.it/dna/ are dedicated to the analysis

of user-submitted DNA sequences; plot.it creates parametric plots of 45

physicochemical, as well as statistical, parameters; bend.it calculates DNA

curvature according to various methods. Both programs provide 1D as well as 2D

plots that allow localisation of peculiar segments within the query. The server

model.it creates 3D models of canonical or bent DNA starting from sequence data

and presents the results in the form of a standard PDB file, directly viewable on

the user's PC using any molecule manipulation program. The recently established

introns server allows statistical evaluation of introns in various taxonomic

groups and the comparison of taxonomic groups in terms of length, base

composition, intron type etc. The options include the analysis of splice sites

and a probability test for exon-shuffling.

PMCID: PMC168966

PMID: 12824394 [Indexed for MEDLINE]

3466. Nucleic Acids Res. 2003 Jul 1;31(13):3666-8.

Cluster-Buster: Finding dense clusters of motifs in DNA sequences.

Frith MC(1), Li MC, Weng Z.

Author information:

(1)Bioinformatics Program, Boston University, 44 Cummington Street, Boston, MA

02215, USA.

The signals that determine activation and repression of specific genes in

response to appropriate stimuli are one of the most important, but least

understood, types of information encoded in genomic DNA. The nucleotide sequence

patterns, or motifs, preferentially bound by various transcription factors have

been collected in databases. However, these motifs appear to be individually too

short and degenerate to enable detection of functional enhancer and silencer

elements within a large genome. Several groups have proposed that dense clusters

of motifs may diagnose regulatory regions more accurately. Cluster-Buster is the

third incarnation of our software for finding clusters of pre-specified motifs in

DNA sequences. We offer a Cluster-Buster web server at

http://zlab.bu.edu/cluster-buster/.

PMCID: PMC168947

PMID: 12824389 [Indexed for MEDLINE]

3467. Nucleic Acids Res. 2003 Jul 1;31(13):3651-3.

Identification of patterns in biological sequences at the ALGGEN server: PROMO

and MALGEN.

Farré D(1), Roset R, Huerta M, Adsuara JE, Roselló L, Albà MM, Messeguer X.

Author information:

(1)Computing Unit, Institut de Recerca Oncològica, L'Hospitalet, Spain.

In this paper we present several web-based tools to identify conserved patterns

in sequences. In particular we present details on the functionality of PROMO

version 2.0, a program for the prediction of transcription factor binding site in

a single sequence or in a group of related sequences and, of MALGEN, a tool to

visualize sequence correspondences among long DNA sequences. The web tools and

associated documentation can be accessed at http://www.lsi.upc.es/~alggen

(RESEARCH link).

PMCID: PMC169011

PMID: 12824386 [Indexed for MEDLINE]

3468. Nucleic Acids Res. 2003 Jul 1;31(13):3645-50.

The web server of IBM's Bioinformatics and Pattern Discovery group.

Huynh T(1), Rigoutsos I, Parida L, Platt D, Shibuya T.

Author information:

(1)Bioinformatics and Pattern Discovery Group, IBM TJ Watson Research Center, PO

BOX 218, Yorktown Heights, NY 10598, USA. rigoutso@us.ibm.com

We herein present and discuss the services and content which are available on the

web server of IBM's Bioinformatics and Pattern Discovery group. The server is

operational around the clock and provides access to a variety of methods that

have been published by the group's members and collaborators. The available tools

correspond to applications ranging from the discovery of patterns in streams of

events and the computation of multiple sequence alignments, to the discovery of

genes in nucleic acid sequences and the interactive annotation of amino acid

sequences. Additionally, annotations for more than 70 archaeal, bacterial,

eukaryotic and viral genomes are available on-line and can be searched

interactively. The tools and code bundles can be accessed beginning at

http://cbcsrv.watson.ibm.com/Tspd.html whereas the genomics annotations are

available at http://cbcsrv.watson.ibm.com/Annotations/.

PMCID: PMC169027

PMID: 12824385 [Indexed for MEDLINE]

3469. Nucleic Acids Res. 2003 Jul 1;31(13):3642-4.

Static benchmarking of membrane helix predictions.

Kernytsky A(1), Rost B.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 650 West 168th Street BB217, New York, NY 10032, USA.

kernytsky@cubic.bioc.columbia.edu

Prediction of trans-membrane helices continues to be a difficult task with a few

prediction methods clearly taking the lead; none of these is clearly best on all

accounts. Recently, we have carefully set up protocols for benchmarking the most

relevant aspects of prediction accuracy and have applied it to >30 prediction

methods. Here, we present the extension of that analysis to the level of an

automatic web server evaluating new methods

(http://cubic.bioc.columbia.edu/services/tmh\_benchmark/). The most important

achievements of the tool are: (i) any new method is compared to the battery of

well-established tools; (ii) the battery of measures explored allows spotting

strengths in methods that may not be 'best' overall. In particular, we report

per-residue and per-segment scores for accuracy and the error-rates for confusing

membrane helices with globular proteins or signal peptides. An additional feature

is that developers can directly investigate any hydrophobicity scale for its

potential in predicting membrane helices.

PMCID: PMC168939

PMID: 12824384 [Indexed for MEDLINE]

3470. Nucleic Acids Res. 2003 Jul 1;31(13):3625-30.

ELM server: A new resource for investigating short functional sites in modular

eukaryotic proteins.

Puntervoll P(1), Linding R, Gemünd C, Chabanis-Davidson S, Mattingsdal M, Cameron

S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferrè F, Maselli V, Via A,

Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli

R, Letunic I, Bork P, Rychlewski L, Küster B, Helmer-Citterich M, Hunter WN,

Aasland R, Gibson TJ.

Author information:

(1)Department of Molecular Biology, University of Bergen, Norway.

Multidomain proteins predominate in eukaryotic proteomes. Individual functions

assigned to different sequence segments combine to create a complex function for

the whole protein. While on-line resources are available for revealing globular

domains in sequences, there has hitherto been no comprehensive collection of

small functional sites/motifs comparable to the globular domain resources, yet

these are as important for the function of multidomain proteins. Short linear

peptide motifs are used for cell compartment targeting, protein-protein

interaction, regulation by phosphorylation, acetylation, glycosylation and a host

of other post-translational modifications. ELM, the Eukaryotic Linear Motif

server at http://elm.eu.org/, is a new bioinformatics resource for investigating

candidate short non-globular functional motifs in eukaryotic proteins, aiming to

fill the void in bioinformatics tools. Sequence comparisons with short motifs are

difficult to evaluate because the usual significance assessments are

inappropriate. Therefore the server is implemented with several logical filters

to eliminate false positives. Current filters are for cell compartment, globular

domain clash and taxonomic range. In favourable cases, the filters can reduce the

number of retained matches by an order of magnitude or more.

PMCID: PMC168952

PMID: 12824381 [Indexed for MEDLINE]

3471. Nucleic Acids Res. 2003 Jul 1;31(13):3621-4.

MHCPred: A server for quantitative prediction of peptide-MHC binding.

Guan P(1), Doytchinova IA, Zygouri C, Flower DR.

Author information:

(1)Edward Jenner Institute for Vaccine Research, High Street, Compton, Berkshire

RG0 7NN, UK.

Accurate T-cell epitope prediction is a principal objective of computational

vaccinology. As a service to the immunology and vaccinology communities at large,

we have implemented, as a server on the World Wide Web, a partial least

squares-based multivariate statistical approach to the quantitative prediction of

peptide binding to major histocom- patibility complexes (MHC), the key checkpoint

on the antigen presentation pathway within adaptive cellular immunity. MHCPred

implements robust statistical models for both Class I alleles (HLA-A\*0101,

HLA-A\*0201, HLA-A\*0202, HLA-A\*0203, HLA-A\*0206, HLA-A\*0301, HLA-A\*1101,

HLA-A\*3301, HLA-A\*6801, HLA-A\*6802 and HLA-B\*3501) and Class II alleles

(HLA-DRB\*0401, HLA-DRB\*0401 and HLA-DRB\*0701). MHCPred is available from the URL:

http://www.jenner.ac.uk/MHCPred.

PMCID: PMC168917

PMID: 12824380 [Indexed for MEDLINE]

3472. Nucleic Acids Res. 2003 Jul 1;31(13):3618-20.

Signal search analysis server.

Ambrosini G(1), Praz V, Jagannathan V, Bucher P.

Author information:

(1)ISREC Swiss Institute for Experimental Cancer Research, Ch. des Boveresses

155, 1066 Epalinges s/ Lausanne, VD, Switzerland.

Signal search analysis is a general method to discover and characterize sequence

motifs that are positionally correlated with a functional site (e.g. a

transcription or translation start site). The method has played an instrumental

role in the analysis of eukaryotic promoter elements. The signal search analysis

server provides access to four different computer programs as well as to a large

number of precompiled functional site collections. The programs offered allow:

(i) the identification of non-random sequence regions under evolutionary

constraint; (ii) the detection of consensus sequence-based motifs that are over-

or under-represented at a particular distance from a functional site; (iii) the

analysis of the positional distribution of a consensus sequence- or weight

matrix-based sequence motif around a functional site; and (iv) the optimization

of a weight matrix description of a locally over-represented sequence motif.

These programs can be accessed at: http://www.isrec.isb-sib.ch/ssa/.

PMCID: PMC169017

PMID: 12824379 [Indexed for MEDLINE]

3473. Nucleic Acids Res. 2003 Jul 1;31(13):3597-600.

GeneSeqer@PlantGDB: Gene structure prediction in plant genomes.

Schlueter SD(1), Dong Q, Brendel V.

Author information:

(1)Department of Zoology and Genetics, Iowa State University, Ames, IA

50011-3260, USA.

The GeneSeqer@PlantGDB Web server (http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi)

provides a gene structure prediction tool tailored for applications to plant

genomic sequences. Predictions are based on spliced alignment with source-native

ESTs and full-length cDNAs or non-native probes derived from putative homologous

genes. The tool is illustrated with applications to refinement of current gene

structure annotation and de novo annotation of draft genomic sequences. The

service should facilitate expert annotation as a community effort by providing

convenient access to all public plant sequences via the PlantGDB database, a

simple four-step protocol for spliced alignment and visually appealing displays

of the predicted gene structures in addition to detailed sequence alignments.

PMCID: PMC168940

PMID: 12824374 [Indexed for MEDLINE]

3474. Nucleic Acids Res. 2003 Jul 1;31(13):3537-9.

RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences.

Wernersson R(1), Pedersen AG.

Author information:

(1)Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University

of Denmark, Building 208, DK-2800, Lyngby, Denmark.

The simple fact that proteins are built from 20 amino acids while DNA only

contains four different bases, means that the 'signal-to-noise ratio' in protein

sequence alignments is much better than in alignments of DNA. Besides this

information-theoretical advantage, protein alignments also benefit from the

information that is implicit in empirical substitution matrices such as

BLOSUM-62. Taken together with the generally higher rate of synonymous mutations

over non-synonymous ones, this means that the phylogenetic signal disappears much

more rapidly from DNA sequences than from the encoded proteins. It is therefore

preferable to align coding DNA at the amino acid level and it is for this purpose

we have constructed the program RevTrans. RevTrans constructs a multiple DNA

alignment by: (i) translating the DNA; (ii) aligning the resulting peptide

sequences; and (iii) building a multiple DNA alignment by 'reverse translation'

of the aligned protein sequences. In the resulting DNA alignment, gaps occur in

groups of three corresponding to entire codons, and analogous codon positions are

therefore always lined up. These features are useful when constructing multiple

DNA alignments for phylogenetic analysis. RevTrans also accepts user-provided

protein alignments for greater control of the alignment process. The RevTrans web

server is freely available at http://www.cbs.dtu.dk/services/RevTrans/.

PMCID: PMC169015

PMID: 12824361 [Indexed for MEDLINE]

3475. Nucleic Acids Res. 2003 Jul 1;31(13):3533-6.

MGAlignIt: A web service for the alignment of mRNA/EST and genomic sequences.

Lee BT(1), Tan TW, Ranganathan S.

Author information:

(1)Department of Biochemistry, National University of Singapore, 8 Medical Drive,

Singapore 117597, Singapore.

Splicing is a biological phenomenon that removes the non-coding sequence from the

transcripts to produce a mature transcript suitable for translation. To study

this phenomenon, information on the intron-exon arrangement of a gene is

essential, usually obtained by aligning mRNA/EST sequences to their cognate

genomic sequences. MGAlign is a novel, rapid, memory efficient and practical

method for aligning mRNA/EST and genome sequences. We present here a freely

available web service, MGAlignIt

(http://origin.bic.nus.edu.sg/mgalign/mgalignit), based on MGAlign. Besides the

alignment itself, this web service allows users to effectively visualize the

alignment in a graphical manner and to perform limited analysis on the alignment

output. The server also permits the alignment to be saved in several forms, both

graphical and text, suitable for further processing and analysis by other

programs.

PMCID: PMC168968

PMID: 12824360 [Indexed for MEDLINE]

3476. Nucleic Acids Res. 2003 Jul 1;31(13):3525-6.

MAVID multiple alignment server.

Bray N(1), Pachter L.

Author information:

(1)Department of Mathematics, 970 Evans Hall, UC Berkeley, Berkeley, CA 94720,

USA.

MAVID is a multiple alignment program suitable for many large genomic regions.

The MAVID web server allows biomedical researchers to quickly obtain multiple

alignments for genomic sequences and to subsequently analyse the alignments for

conserved regions. MAVID has been successfully used for the alignment of closely

related species such as primates and also for the alignment of more distant

organisms such as human and fugu. The server is fast, capable of aligning

hundreds of kilobases in less than a minute. The multiple alignment is used to

build a phylogenetic tree for the sequences, which is subsequently used as a

basis for identifying conserved regions in the alignment. The server can be

accessed at http://baboon.math.berkeley.edu/mavid/.

PMCID: PMC169029

PMID: 12824358 [Indexed for MEDLINE]

3477. Nucleic Acids Res. 2003 Jul 1;31(13):3518-24.

MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic

DNA sequences.

Schwartz S(1), Elnitski L, Li M, Weirauch M, Riemer C, Smit A; NISC Comparative

Sequencing Program, Green ED, Hardison RC, Miller W.

Author information:

(1)Department of Computer Science and Engineering, The Pennsylvania State

University, University Park, PA 16802, USA.

Analysis of multiple sequence alignments can generate important, testable

hypotheses about the phylogenetic history and cellular function of genomic

sequences. We describe the MultiPipMaker server, which aligns multiple, long

genomic DNA sequences quickly and with good sensitivity (available at

http://bio.cse.psu.edu/ since May 2001). Alignments are computed between a

contiguous reference sequence and one or more secondary sequences, which can be

finished or draft sequence. The outputs include a stacked set of percent identity

plots, called a MultiPip, comparing the reference sequence with subsequent

sequences, and a nucleotide-level multiple alignment. New tools are provided to

search MultiPipMaker output for conserved matches to a user-specified pattern and

for conserved matches to position weight matrices that describe transcription

factor binding sites (singly and in clusters). We illustrate the use of

MultiPipMaker to identify candidate regulatory regions in WNT2 and then

demonstrate by transfection assays that they are functional. Analysis of the

alignments also confirms the phylogenetic inference that horses are more closely

related to cats than to cows.

PMCID: PMC168985

PMID: 12824357 [Indexed for MEDLINE]

3478. Nucleic Acids Res. 2003 Jul 1;31(13):3507-9.

SLAM web server for comparative gene finding and alignment.

Cawley S(1), Pachter L, Alexandersson M.

Author information:

(1)Affymetrix Inc., 6550 Vallejo St, Suite 100, Emeryville, CA 94608, USA.

simon\_cawley@affymetrix.com

SLAM is a program that simultaneously aligns and annotates pairs of homologous

sequences. The SLAM web server integrates SLAM with repeat masking tools and the

AVID alignment program to allow for rapid alignment and gene prediction in user

submitted sequences. Along with annotations and alignments for the submitted

sequences, users obtain a list of predicted conserved non-coding sequences (and

their associated alignments). The web site also links to whole genome annotations

of the human, mouse and rat genomes produced with the SLAM program. The server

can be accessed at http://bio.math.berkeley.edu/slam.

PMCID: PMC168989

PMID: 12824355 [Indexed for MEDLINE]

3479. Nucleic Acids Res. 2003 Jul 1;31(13):3503-6.

Tcoffee@igs: A web server for computing, evaluating and combining multiple

sequence alignments.

Poirot O(1), O'Toole E, Notredame C.

Author information:

(1)Information Genomique et Structurale, CNRS, 31 Chemin Joseph Aiguier, 13 402

Marseille Cedex 20, France.

This paper presents Tcoffee@igs, a new server provided to the community by Hewlet

Packard computers and the Centre National de la Recherche Scientifique. This

server is a web-based tool dedicated to the computation, the evaluation and the

combination of multiple sequence alignments. It uses the latest version of the

T-Coffee package. Given a set of unaligned sequences, the server returns an

evaluated multiple sequence alignment and the associated phylogenetic tree. This

server also makes it possible to evaluate the local reliability of an existing

alignment and to combine several alternative multiple alignments into a single

new one. Tcoffee@igs can be used for aligning protein, RNA or DNA sequences.

Datasets of up to 100 sequences (2000 residues long) can be processed. The server

and its documentation are available from: http://igs-server.cnrs-mrs.fr/Tcoffee/.

PMCID: PMC168929

PMID: 12824354 [Indexed for MEDLINE]

3480. Nucleic Acids Res. 2003 Jul 1;31(13):3501-2.

CLOURE: Clustal Output Reformatter, a program for reformatting ClustalX/ClustalW

outputs for SNP analysis and molecular systematics.

Kohli DK(1), Bachhawat AK.

Author information:

(1)Institute of Microbial Technology, Sector 39-A, Chandigarh 160 036, India.

We describe a program (and a website) to reformat the ClustalX/ClustalW outputs

to a format that is widely used in the presentation of sequence alignment data in

SNP analysis and molecular systematic studies. This program, CLOURE, CLustal

OUtput REformatter, takes the multiple sequence alignment file (nucleic acid or

protein) generated from Clustal as input files. The CLOURE-D format presents the

Clustal alignment in a format that highlights only the different

nucleotides/residues relative to the first query sequence. The program has been

written in Visual Basic and will run on a Windows platform. The downloadable

program, as well as a web-based server which has also been developed, can be

accessed at http://imtech.res.in/~anand/cloure.html.

PMCID: PMC168909

PMID: 12824353 [Indexed for MEDLINE]

3481. Nucleic Acids Res. 2003 Jul 1;31(13):3487-90.

REDUCE: An online tool for inferring cis-regulatory elements and transcriptional

module activities from microarray data.

Roven C(1), Bussemaker HJ.

Author information:

(1)Department of Biological Sciences, Columbia University, New York, NY, USA.

REDUCE is a motif-based regression method for microarray analysis. The only

required inputs are (i) a single genome-wide set of absolute or relative mRNA

abundances and (ii) the DNA sequence of the regulatory region associated with

each gene that is probed. Currently supported organisms are yeast, worm and fly;

it is an open question whether in its current incarnation our approach can be

used for mouse or human. REDUCE uses unbiased statistics to identify

oligonucleotide motifs whose occurrence in the regulatory region of a gene

correlates with the level of mRNA expression. Regression analysis is used to

infer the activity of the transcriptional module associated with each motif.

REDUCE is available online at http://bussemaker.bio.columbia.edu/reduce/. This

web site provides functionality for the upload and management of microarray data.

REDUCE analysis results can be viewed and downloaded, and optionally be shared

with other users or made publicly accessible.

PMCID: PMC169192

PMID: 12824350 [Indexed for MEDLINE]

3482. Nucleic Acids Res. 2003 Jul 1;31(13):3471-6.

GenePublisher: Automated analysis of DNA microarray data.

Knudsen S(1), Workman C, Sicheritz-Ponten T, Friis C.

Author information:

(1)Center for Biological Sequence Analysis, BioCentrum-DTU, 2800 Lyngby, Denmark.

steen@cbs.dtu.dk

GenePublisher, a system for automatic analysis of data from DNA microarray

experiments, has been implemented with a web interface at

http://www.cbs.dtu.dk/services/GenePublisher. Raw data are uploaded to the server

together with a specification of the data. The server performs normalization,

statistical analysis and visualization of the data. The results are run against

databases of signal transduction pathways, metabolic pathways and promoter

sequences in order to extract more information. The results of the entire

analysis are summarized in report form and returned to the user.

PMCID: PMC169191

PMID: 12824347 [Indexed for MEDLINE]

3483. Nucleic Acids Res. 2003 Jul 1;31(13):3461-7.

GEPAS: A web-based resource for microarray gene expression data analysis.

Herrero J(1), Al-Shahrour F, Díaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo J,

Dopazo J.

Author information:

(1)Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas, c/Melchor

Fernández Almagro 3, 28029, Madrid, Spain.

We present a web-based pipeline for microarray gene expression profile analysis,

GEPAS, which stands for Gene Expression Profile Analysis Suite

(http://gepas.bioinfo.cnio.es). GEPAS is composed of different interconnected

modules which include tools for data pre-processing, two-conditions comparison,

unsupervised and supervised clustering (which include some of the most popular

methods as well as home made algorithms) and several tests for differential gene

expression among different classes, continuous variables or survival analysis. A

multiple purpose tool for data mining, based on Gene Ontology, is also linked to

the tools, which constitutes a very convenient way of analysing clustering

results. On-line tutorials are available from our main web server

(http://bioinfo.cnio.es).

PMCID: PMC168997

PMID: 12824345 [Indexed for MEDLINE]

3484. Nucleic Acids Res. 2003 Jul 1;31(13):3450-60.

Tools for the automatic identification and classification of RNA base pairs.

Yang H(1), Jossinet F, Leontis N, Chen L, Westbrook J, Berman H, Westhof E.

Author information:

(1)Department of Chemistry and Chemical Biology, Rutgers University, NJ

08854-8087, USA.

Three programs have been developed to aid in the classification and visualization

of RNA structure. BPViewer provides a web interface for displaying

three-dimensional (3D) coordinates of individual base pairs or base pair

collections. A web server, RNAview, automatically identifies and classifies the

types of base pairs that are formed in nucleic acid structures by various

combinations of the three edges, Watson-Crick, Hoogsteen and the Sugar edge.

RNAView produces two-dimensional (2D) diagrams of secondary and tertiary

structure in either Postscript, VRML or RNAML formats. The application RNAMLview

can be used to rearrange various parts of the RNAView 2D diagram to generate a

standard representation (like the cloverleaf structure of tRNAs) or any layout

desired by the user. A 2D diagram can be rapidly reformatted using RNAMLview

since all the parts of RNA (like helices and single strands) are dynamically

linked while moving the selected parts. With the base pair annotation and the 2D

graphic display, RNA motifs are rapidly identified and classified. A survey has

been carried out for 41 unique structures selected from the NDB database. The

statistics for the occurrence of each edge and of each of the 12 bp families are

given for the combinations of the four bases: A, G, U and C. The program also

allows for visualization of the base pair interactions by using a symbolic

convention previously proposed for base pairs. The web servers for BPViewer and

RNAview are available at http://ndbserver.rutgers.edu/services/. The application

RNAMLview can also be downloaded from this site. The 2D diagrams produced by

RNAview are available for RNA structures in the Nucleic Acid Database (NDB) at

http://ndbserver.rutgers.edu/atlas/.

PMCID: PMC168936

PMID: 12824344 [Indexed for MEDLINE]

3485. Nucleic Acids Res. 2003 Jul 1;31(13):3441-5.

A software tool-box for analysis of regulatory RNA elements.

Bengert P(1), Dandekar T.

Author information:

(1)Parasitology, University of Heidelberg, Im Neuenheimer Feld 324, 69120

Heidelberg, Germany.

We describe an integrated tool-box to identify regulatory RNA elements. The RNA

analyzer collects general and specific information on any submitted RNA sequence

or batch of sequences in FASTA format. It determines and rapidly scans the

different regions of an RNA (including 5' UTR, CDS, 3' UTR in mRNA) and screens

for specific RNA signals (in each of these regions, e.g. polyA-site, AU rich

region etc. in 3' UTR). It runs a fast folding RNA routine to provide an overview

of the RNA fold. Furthermore it analyzes structure content, fold energy and stem

loops. In addition, consensus templates are used to determine whether there are

any functional structures present for translational control (template: IRE),

structured RNA (template: tRNA consensus) or catalytic RNA (template:

trans-splicing RNA), giving indications as to how well the structures found match

to these templates. The tool box has been implemented as a WWW server at

http://wb2x01.biozentrum.uni-wuerzburg.de/.

PMCID: PMC168974

PMID: 12824342 [Indexed for MEDLINE]

3486. Nucleic Acids Res. 2003 Jul 1;31(13):3429-31.

Vienna RNA secondary structure server.

Hofacker IL(1).

Author information:

(1)Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität

Wien, Währingerstrasse 17, A-1090 Wien, Austria. ivo@tbi.univie.ac.at

The Vienna RNA secondary structure server provides a web interface to the most

frequently used functions of the Vienna RNA software package for the analysis of

RNA secondary structures. It currently offers prediction of secondary structure

from a single sequence, prediction of the consensus secondary structure for a set

of aligned sequences and the design of sequences that will fold into a predefined

structure. All three services can be accessed via the Vienna RNA web server at

http://rna.tbi.univie.ac.at/.

PMCID: PMC169005

PMID: 12824340 [Indexed for MEDLINE]

3487. Nucleic Acids Res. 2003 Jul 1;31(13):3423-8.

Pfold: RNA secondary structure prediction using stochastic context-free grammars.

Knudsen B(1), Hein J.

Author information:

(1)BiRC (Bioinformatics Research Center), University of Aarhus, Høegh

Guldbergsgade 10, Building 090, 8000 Arhus C, Denmark. bk@birc.dk

RNA secondary structures are important in many biological processes and efficient

structure prediction can give vital directions for experimental investigations.

Many available programs for RNA secondary structure prediction only use a single

sequence at a time. This may be sufficient in some applications, but often it is

possible to obtain related RNA sequences with conserved secondary structure.

These should be included in structural analyses to give improved results. This

work presents a practical way of predicting RNA secondary structure that is

especially useful when related sequences can be obtained. The method improves a

previous algorithm based on an explicit evolutionary model and a probabilistic

model of structures. Predictions can be done on a web server at

http://www.daimi.au.dk/~compbio/pfold.

PMCID: PMC169020

PMID: 12824339 [Indexed for MEDLINE]

3488. Nucleic Acids Res. 2003 Jul 1;31(13):3406-15.

Mfold web server for nucleic acid folding and hybridization prediction.

Zuker M(1).

Author information:

(1)Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy,

NY 12180, USA. zukerm@rpi.edu

The abbreviated name, 'mfold web server', describes a number of closely related

software applications available on the World Wide Web (WWW) for the prediction of

the secondary structure of single stranded nucleic acids. The objective of this

web server is to provide easy access to RNA and DNA folding and hybridization

software to the scientific community at large. By making use of universally

available web GUIs (Graphical User Interfaces), the server circumvents the

problem of portability of this software. Detailed output, in the form of

structure plots with or without reliability information, single strand frequency

plots and 'energy dot plots', are available for the folding of single sequences.

A variety of 'bulk' servers give less information, but in a shorter time and for

up to hundreds of sequences at once. The portal for the mfold web server is

http://www.bioinfo.rpi.edu/applications/mfold. This URL will be referred to as

'MFOLDROOT'.

PMCID: PMC169194

PMID: 12824337 [Indexed for MEDLINE]

3489. Nucleic Acids Res. 2003 Jul 1;31(13):3404-5.

SSEP: Secondary structural elements of proteins.

Shanthi V(1), Selvarani P, Kumar ChK, Mohire CS, Sekar K.

Author information:

(1)Bioinformatics Centre, Supercomputer Education and Research Centre, Indian

Institute of Science, Bangalore 560 012, India.

SSEP is a comprehensive resource for accessing information related to the

secondary structural elements present in the 25 and 90% non-redundant protein

chains. The database contains 1771 protein chains from 1670 protein structures

and 6182 protein chains from 5425 protein structures in 25 and 90% non-redundant

protein chains, respectively. The current version provides information about the

alpha-helical segments and beta-strand fragments of varying lengths. In addition,

it also contains the information about 3(10)-helix, beta- and nu-turns and

hairpin loops. The free graphics program RASMOL has been interfaced with the

search engine to visualize the three-dimensional structures of the user queried

secondary structural fragment. The database is updated regularly and is available

through Bioinformatics web server at http://cluster.physics.iisc.ernet.in/ssep/

or http://144.16.71.148/ssep/.

PMCID: PMC168914

PMID: 12824336 [Indexed for MEDLINE]

3490. Nucleic Acids Res. 2003 Jul 1;31(13):3393-9.

Integrated databanks access and sequence/structure analysis services at the PBIL.

Perrière G(1), Combet C, Penel S, Blanchet C, Thioulouse J, Geourjon C, Grassot

J, Charavay C, Gouy M, Duret L, Deléage G.

Author information:

(1)Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS no. 5558, Université

Claude Bernard, Lyon 1, 43 bd du 11 Novembre 1918, 69622 Villeurbanne Cedex,

France. perriere@biomserv.univ-lyon1.fr

The World Wide Web server of the PBIL (Pôle Bioinformatique Lyonnais) provides

on-line access to sequence databanks and to many tools of nucleic acid and

protein sequence analyses. This server allows to query nucleotide sequence banks

in the EMBL and GenBank formats and protein sequence banks in the SWISS-PROT and

PIR formats. The query engine on which our data bank access is based is the ACNUC

system. It allows the possibility to build complex queries to access functional

zones of biological interest and to retrieve large sequence sets. Of special

interest are the unique features provided by this system to query the data banks

of gene families developed at the PBIL. The server also provides access to a wide

range of sequence analysis methods: similarity search programs, multiple

alignments, protein structure prediction and multivariate statistics. An

originality of this server is the integration of these two aspects: sequence

retrieval and sequence analysis. Indeed, thanks to the introduction of re-usable

lists, it is possible to perform treatments on large sets of data. The PBIL

server can be reached at: http://pbil.univ-lyon1.fr.

PMCID: PMC168937

PMID: 12824334 [Indexed for MEDLINE]

3491. Nucleic Acids Res. 2003 Jul 1;31(13):3381-5.

SWISS-MODEL: An automated protein homology-modeling server.

Schwede T(1), Kopp J, Guex N, Peitsch MC.

Author information:

(1)Biozentrum der Universität Basel, Klingelbergstr. 50-70, CH 4056 Basel,

Switzerland. torsten.schwede@unibas.ch

SWISS-MODEL (http://swissmodel.expasy.org) is a server for automated comparative

modeling of three-dimensional (3D) protein structures. It pioneered the field of

automated modeling starting in 1993 and is the most widely-used free web-based

automated modeling facility today. In 2002 the server computed 120 000 user

requests for 3D protein models. SWISS-MODEL provides several levels of user

interaction through its World Wide Web interface: in the 'first approach mode'

only an amino acid sequence of a protein is submitted to build a 3D model.

Template selection, alignment and model building are done completely automated by

the server. In the 'alignment mode', the modeling process is based on a

user-defined target-template alignment. Complex modeling tasks can be handled

with the 'project mode' using DeepView (Swiss-PdbViewer), an integrated

sequence-to-structure workbench. All models are sent back via email with a

detailed modeling report. WhatCheck analyses and ANOLEA evaluations are provided

optionally. The reliability of SWISS-MODEL is continuously evaluated in the

EVA-CM project. The SWISS-MODEL server is under constant development to improve

the successful implementation of expert knowledge into an easy-to-use server.

PMCID: PMC168927

PMID: 12824332 [Indexed for MEDLINE]

3492. Nucleic Acids Res. 2003 Jul 1;31(13):3375-80.

Tools for comparative protein structure modeling and analysis.

Eswar N(1), John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC,

Marti-Renom MA, Madhusudhan MS, Yerkovich B, Sali A.

Author information:

(1)Department of Biopharmaceutical Sciences and California Institute for

Quantitative Biomedical Research, University of California, San Francisco, CA

94143-2240, USA.

The following resources for comparative protein structure modeling and analysis

are described (http://salilab.org): MODELLER, a program for comparative modeling

by satisfaction of spatial restraints; MODWEB, a web server for automated

comparative modeling that relies on PSI-BLAST, IMPALA and MODELLER; MODLOOP, a

web server for automated loop modeling that relies on MODELLER; MOULDER, a CPU

intensive protocol of MODWEB for building comparative models based on distant

known structures; MODBASE, a comprehensive database of annotated comparative

models for all sequences detectably related to a known structure; MODVIEW, a

Netscape plugin for Linux that integrates viewing of multiple sequences and

structures; and SNPWEB, a web server for structure-based prediction of the

functional impact of a single amino acid substitution.

PMCID: PMC168950

PMID: 12824331 [Indexed for MEDLINE]

3493. Nucleic Acids Res. 2003 Jul 1;31(13):3367-9.

MATRAS: A program for protein 3D structure comparison.

Kawabata T(1).

Author information:

(1)Graduate School of Information Science, Nara Institute of Science and

Technology, Takayama 8916-5, Ikoma, Nara, Japan. takawaba@is.aist-nara.ac.jp

The recent accumulation of large amounts of 3D structural data warrants a

sensitive and automatic method to compare and classify these structures. We

developed a web server for comparing protein 3D structures using the program

Matras (http://biunit.aist-nara.ac.jp/matras). An advantage of Matras is its

structure similarity score, which is defined as the log-odds of the

probabilities, similar to Dayhoff's substitution model of amino acids. This score

is designed to detect evolutionarily related (homologous) structural

similarities. Our web server has three main services. The first one is a pairwise

3D alignment, which is simply align two structures. A user can assign structures

by either inputting PDB codes or by uploading PDB format files in the local

machine. The second service is a multiple 3D alignment, which compares several

protein structures. This program employs the progressive alignment algorithm, in

which pairwise 3D alignments are assembled in the proper order. The third service

is a 3D library search, which compares one query structure against a large number

of library structures. We hope this server provides useful tools for insights

into protein 3D structures.

PMCID: PMC168987

PMID: 12824329 [Indexed for MEDLINE]

3494. Nucleic Acids Res. 2003 Jul 1;31(13):3349-51.

MolSurfer: A macromolecular interface navigator.

Gabdoulline RR(1), Wade RC, Walther D.

Author information:

(1)European Media Laboratory (EML), Schloss-Wolfsbrunnenweg 33, D-69118,

Heidelberg, Germany. razif.gabdoulline@eml.villa-bosch.de

We describe the current status of the Java molecular graphics tool, MolSurfer.

MolSurfer has been designed to assist the analysis of the structures and

physico-chemical properties of macromolecular interfaces. MolSurfer provides a

coupled display of two-dimensional (2D) maps of the interfaces generated with the

ADS software and a three-dimensional (3D) view of the macromolecular structure in

the Java PDB viewer, WebMol. The interfaces are analytically defined and

properties such as electrostatic potential or hydrophobicity are projected on to

them. MolSurfer has been applied previously to analyze a set of 39

protein-protein complexes, with structures available from the Protein Data Bank

(PDB). A new application, described here, is the visualization of 75 interfaces

in structures of protein-DNA and protein-RNA complexes. Another new feature is

that the MolSurfer web server is now able to compute and map Poisson-Boltzmann

electrostatic potentials of macromolecules onto interfaces. The MolSurfer web

server is available at http://projects.villa-bosch.de/mcm/software/molsurfer.

PMCID: PMC168994

PMID: 12824324 [Indexed for MEDLINE]

3495. Nucleic Acids Res. 2003 Jul 1;31(13):3345-8.

NCI: A server to identify non-canonical interactions in protein structures.

Babu MM(1).

Author information:

(1)MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

madanm@mrc-lmb.cam.ac.uk

NCI is a server for the identification of non-canonical interactions in protein

structures. These interactions, which include N-H...pi, C(alpha)-H...pi,

C(alpha)-H...O=C and variants of them, were first observed in small molecules and

subsequently in high-resolution protein structures. Such interactions have been

subjected to extensive structural analysis to elucidate the different geometric

criteria required to identify them. These interactions have also recently been

shown to be important for the stability of protein structures. In this work, I

describe a server called NCI, which allows the user to either upload

protein/peptide coordinates in Protein Data Bank (PDB) format or enter a

Structural Classification of Proteins database (SCOP)/PDB identifier for which

NCI identifies the different non-canonical interactions, based purely on

geometric criteria. Results are presented as an HTML table, as a parseable text

file and as a color-coded interaction matrix. In addition, the user can view the

RasMol image highlighting the interactions in the protein structure and download

the RasMol script. The NCI server is available at:

http://www.mrc-lmb.cam.ac.uk/genomes/nci/.

PMCID: PMC168935

PMID: 12824323 [Indexed for MEDLINE]

3496. Nucleic Acids Res. 2003 Jul 1;31(13):3337-40.

LOC3D: annotate sub-cellular localization for protein structures.

Nair R(1), Rost B.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 650 West 168th Street BB217, New York, NY 10032, USA.

nair@cubic.bioc.columbia.edu

LOC3D (http://cubic.bioc.columbia.edu/db/LOC3d/) is both a weekly-updated

database and a web server for predictions of sub-cellular localization for

eukaryotic proteins of known three-dimensional (3D) structure. Localization is

predicted using four different methods: (i) PredictNLS, prediction of nuclear

proteins through nuclear localization signals; (ii) LOChom, inferring

localization through sequence homology; (iii) LOCkey, inferring localization

through automatic text analysis of SWISS-PROT keywords; and (iv) LOC3Dini, ab

initio prediction through a system of neural networks and vector support

machines. The final prediction is based on the method that predicts localization

with the highest confidence. The LOC3D database currently contains predictions

for >8700 eukaryotic protein chains taken from the Protein Data Bank (PDB). The

web server can be used to predict sub-cellular localization for proteins for

which only a predicted structure is available from threading servers. This makes

the resource of particular interest to structural genomics initiatives.

PMCID: PMC168921

PMID: 12824321 [Indexed for MEDLINE]

3497. Nucleic Acids Res. 2003 Jul 1;31(13):3320-3.

ESPript/ENDscript: Extracting and rendering sequence and 3D information from

atomic structures of proteins.

Gouet P(1), Robert X, Courcelle E.

Author information:

(1)Laboratoire de BioCristallographie, IBCP-CNRS UMR 5086 UCBL, 7 passage du

Vercors, 69367 Lyon Cedex 07, France. p.gouet@ibcp.fr

The fortran program ESPript was created in 1993, to display on a PostScript

figure multiple sequence alignments adorned with secondary structure elements. A

web server was made available in 1999 and ESPript has been linked to three major

web tools: ProDom which identifies protein domains, PredictProtein which predicts

secondary structure elements and NPS@ which runs sequence alignment programs. A

web server named ENDscript was created in 2002 to facilitate the generation of

ESPript figures containing a large amount of information. ENDscript uses programs

such as BLAST, Clustal and PHYLODENDRON to work on protein sequences and such as

DSSP, CNS and MOLSCRIPT to work on protein coordinates. It enables the creation,

from a single Protein Data Bank identifier, of a multiple sequence alignment

figure adorned with secondary structure elements of each sequence of known 3D

structure. Similar 3D structures are superimposed in turn with the program PROFIT

and a final figure is drawn with BOBSCRIPT, which shows sequence and structure

conservation along the Calpha trace of the query. ESPript and ENDscript are

available at http://genopole.toulouse.inra.fr/ESPript.

PMCID: PMC168963

PMID: 12824317 [Indexed for MEDLINE]

3498. Nucleic Acids Res. 2003 Jul 1;31(13):3316-9.

VADAR: a web server for quantitative evaluation of protein structure quality.

Willard L(1), Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, Wishart DS.

Author information:

(1)Department of Biochemistry, University of Alberta, Edmonton, AB, T6G 2N8,

Canada.

VADAR (Volume Area Dihedral Angle Reporter) is a comprehensive web server for

quantitative protein structure evaluation. It accepts Protein Data Bank (PDB)

formatted files or PDB accession numbers as input and calculates, identifies,

graphs, reports and/or evaluates a large number (>30) of key structural

parameters both for individual residues and for the entire protein. These include

excluded volume, accessible surface area, backbone and side chain dihedral

angles, secondary structure, hydrogen bonding partners, hydrogen bond energies,

steric quality, solvation free energy as well as local and overall fold quality.

These derived parameters can be used to rapidly identify both general and

residue-specific problems within newly determined protein structures. The VADAR

web server is freely accessible at http://redpoll.pharmacy.ualberta.ca/vadar.

PMCID: PMC168972

PMID: 12824316 [Indexed for MEDLINE]

3499. Nucleic Acids Res. 2003 Jul 1;31(13):3311-5.

EVA: Evaluation of protein structure prediction servers.

Koh IY(1), Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N,

Graña O, Pazos F, Valencia A, Sali A, Rost B.

Author information:

(1)Columbia University Center for Computational Biology and Bioinformatics

(C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York, NY 10032, USA.

koh@cubic.bioc.columbia.edu

EVA (http://cubic.bioc.columbia.edu/eva/) is a web server for evaluation of the

accuracy of automated protein structure prediction methods. The evaluation is

updated automatically each week, to cope with the large number of existing

prediction servers and the constant changes in the prediction methods. EVA

currently assesses servers for secondary structure prediction, contact

prediction, comparative protein structure modelling and threading/fold

recognition. Every day, sequences of newly available protein structures in the

Protein Data Bank (PDB) are sent to the servers and their predictions are

collected. The predictions are then compared to the experimental structures once

a week; the results are published on the EVA web pages. Over time, EVA has

accumulated prediction results for a large number of proteins, ranging from

hundreds to thousands, depending on the prediction method. This large sample

assures that methods are compared reliably. As a result, EVA provides useful

information to developers as well as users of prediction methods.

PMCID: PMC169025

PMID: 12824315 [Indexed for MEDLINE]

3500. Nucleic Acids Res. 2003 Jul 1;31(13):3308-10.

META-PP: single interface to crucial prediction servers.

Eyrich VA(1), Rost B.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 650 West 168th Street BB217, New York, NY 10032, USA.

volker@chem.columbia.edu

The META-PP server (http://cubic.bioc.columbia.edu/meta/) simplifies access to a

battery of public protein structure and function prediction servers by providing

a common and stable web-based interface. The goal is to make these powerful and

increasingly essential methods more readily available to nonexpert users and the

bioinformatics community at large. At present META-PP provides access to a

selected set of high-quality servers in the areas of comparative modelling,

threading/fold recognition, secondary structure prediction and more specialized

fields like contact and function prediction.

PMCID: PMC168978

PMID: 12824314 [Indexed for MEDLINE]

3501. Nucleic Acids Res. 2003 Jul 1;31(13):3305-7.

GeneSilico protein structure prediction meta-server.

Kurowski MA(1), Bujnicki JM.

Author information:

(1)Bioinformatics Laboratory, International Institute of Molecular and Cell

Biology, Warsaw, Poland.

Rigorous assessments of protein structure prediction have demonstrated that fold

recognition methods can identify remote similarities between proteins when

standard sequence search methods fail. It has been shown that the accuracy of

predictions is improved when refined multiple sequence alignments are used

instead of single sequences and if different methods are combined to generate a

consensus model. There are several meta-servers available that integrate protein

structure predictions performed by various methods, but they do not allow for

submission of user-defined multiple sequence alignments and they seldom offer

confidentiality of the results. We developed a novel WWW gateway for protein

structure prediction, which combines the useful features of other meta-servers

available, but with much greater flexibility of the input. The user may submit an

amino acid sequence or a multiple sequence alignment to a set of methods for

primary, secondary and tertiary structure prediction. Fold-recognition results

(target-template alignments) are converted into full-atom 3D models and the

quality of these models is uniformly assessed. A consensus between different FR

methods is also inferred. The results are conveniently presented on-line on a

single web page over a secure, password-protected connection. The GeneSilico

protein structure prediction meta-server is freely available for academic users

at http://genesilico.pl/meta.

PMCID: PMC168964

PMID: 12824313 [Indexed for MEDLINE]

3502. Nucleic Acids Res. 2003 Jul 1;31(13):3300-4.

The PredictProtein server.

Rost B(1), Liu J.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, 650 West 168th Street BB217, New York, NY 10032, USA.

rost@columbia.edu

PredictProtein (PP, http://cubic.bioc.columbia.edu/pp/) is an internet service

for sequence analysis and the prediction of aspects of protein structure and

function. Users submit protein sequence or alignments; the server returns a

multiple sequence alignment, PROSITE sequence motifs, low-complexity regions

(SEG), ProDom domain assignments, nuclear localisation signals, regions lacking

regular structure and predictions of secondary structure, solvent accessibility,

globular regions, transmembrane helices, coiled-coil regions, structural switch

regions and disulfide-bonds. Upon request, fold recognition by prediction-based

threading is available. For all services, users can submit their query either by

electronic mail or interactively from World Wide Web.

PMCID: PMC168915

PMID: 12824312 [Indexed for MEDLINE]

3503. Nucleic Acids Res. 2003 Jul 1;31(13):3296-9.

PROTINFO: Secondary and tertiary protein structure prediction.

Hung LH(1), Samudrala R.

Author information:

(1)Computational Genomics Group, Department of Microbiology, University of

Washington School of Medicine, Seattle, WA 98195, USA.

Information about the secondary and tertiary structure of a protein sequence can

greatly assist biologists in the generation and testing of hypotheses, as well as

design of experiments. The PROTINFO server enables users to submit a protein

sequence and request a prediction of the three-dimensional (tertiary) structure

based on comparative modeling, fold generation and de novo methods developed by

the authors. In addition, users can submit NMR chemical shift data and request

protein secondary structure assignment that is based on using neural networks to

combine the chemical shifts with secondary structure predictions. The server is

available at http://protinfo.compbio.washington.edu.

PMCID: PMC168948

PMID: 12824311 [Indexed for MEDLINE]

3504. Nucleic Acids Res. 2003 Jul 1;31(13):3291-2.

Detection of reliable and unexpected protein fold predictions using 3D-Jury.

Ginalski K(1), Rychlewski L.

Author information:

(1)Bioinformatics Laboratory, BioInfoBank Institute, ul. Limanowskiego 24A,

60-744 Poznan, Poland.

3D-Jury is a fully automated protein structure meta prediction system accessible

via the Meta Server interface (http://BioInfo.PL/Meta). This is one of the meta

predictors, which have made a dramatic, unprecedented impact on the last CASP-5

experiment. The 3D-Jury is comparable with other meta servers but it has the

highest combined specificity and sensitivity. The presented method is also very

simple and versatile and can be used to create meta predictions even from sets of

models produced by humans. An additional and very important and novel feature of

the system is the high correlation between the reported confidence score and the

accuracy of the model. The number of correctly predicted residues can be

estimated directly from the prediction score. The high reliability of the method

enables any biologist to submit a target of interest to the Meta Server and

screen with relatively high confidence, whether the target can be predicted by

fold recognition methods while being unpredictable using standard approaches like

PSI-Blast. This can point to interesting relationships which could have been

missed in annotations of proteins or genomes and provide very valuable

information for novel scientific discoveries.

PMCID: PMC168910

PMID: 12824309 [Indexed for MEDLINE]

3505. Protein Sci. 2003 Jul;12(7):1547-55.

Predicting the topology of transmembrane helical proteins using mean burial

propensity and a hidden-Markov-model-based method.

Zhou H(1), Zhou Y.

Author information:

(1)Center for Single Molecule Biophysics, Department of Physiology & Biophysics,

Howard Hughes Medical Institute, State University of New York at Buffalo,

Buffalo, NY 14214, USA.

Helices in membrane spanning regions are more tightly packed than the helices in

soluble proteins. Thus, we introduce a method that uses a simple scale of burial

propensity and a new algorithm to predict transmembrane helical (TMH) segments

and a positive-inside rule to predict amino-terminal orientation. The method (the

topology predictor of transmembrane helical proteins using mean burial propensity

[THUMBUP]) correctly predicted the topology of 55 of 73 proteins (or 75%) with

known three-dimensional structures (the 3D helix database). This level of

accuracy can be reached by MEMSAT 1.8 (a 200-parameter model-recognition method)

and a new HMM-based method (a 111-parameter hidden Markov model, UMDHMM(TMHP)) if

they were retrained with the 73-protein database. Thus, a method based on a

physiochemical property can provide topology prediction as accurate as those

methods based on more complicated statistical models and learning algorithms for

the proteins with accurately known structures. Commonly used HMM-based methods

and MEMSAT 1.8 were trained with a combination of the partial 3D helix database

and a 1D helix database of TMH proteins in which topology information were

obtained by gene fusion and other experimental techniques. These methods provide

a significantly poorer prediction for the topology of TMH proteins in the 3D

helix database. This suggests that the 1D helix database, because of its

inaccuracy, should be avoided as either a training or testing database. A Web

server of THUMBUP and UMDHMM(TMHP) is established for academic users at

http://www.smbs.buffalo.edu/phys\_bio/service.htm. The 3D helix database is also

available from the same Web site.

DOI: 10.1110/ps.0305103

PMCID: PMC2323935

PMID: 12824500 [Indexed for MEDLINE]

3506. Sheng Wu Gong Cheng Xue Bao. 2003 Jul;19(4):493-6.

[Studies on the mechanism of thermostability and thermophilicity change of

thermostable alkaline phosphatase and its mutants].

[Article in Chinese]

Yu F(1), Xu XF, Jin Z.

Author information:

(1)College of Life Sciences, Nanjing Normal University, Nanjing 210097, China.

The relationship among the substituted amino acids, the 3D structure simulated on

PC through CPHmodels Server ( http://www.cbs.dtu. dk/services/CPHmodels/) and the

thermostable performance of 4 thermostable alkaline phosphatase(TAP) mutants

selected from a clone bank of more than 200 mutants were analyzed to explore the

mechanism of thermostability change. These mutants are TAP(A410T) (A410-->T),

TAP(P396S) (P396-->S), TAP2(N100S T320-->I) and TAP4(N100-->S P396-->S A410 -->V

P490-->S). TAP and the mutants' thermostable performance was evaluated by

measuring the highest tolerable temperature (T1/2) and the optimal reaction

temperature (Topt). The 3D structure neighboring the substituted amino acids was

simulated by Swiss-PDBViewer to observe the relationship between the structure

change and the thermostable performance of TAP and its mutants. The results

displayed that all these amino acid substitutions except the T320-->I mutant

brought about only a little local change on TAP's 3D structure and very little

effect on their optimal reaction temperature, but a significant decrease (nearly

10 degrees C) on their highest tolerable temperature. However, the T320-->I

mutation due to close to TAP's active sites did bring about a significant

descendents of the mutant in both the highest tolerable temperature and the

optimal reaction temperature. Thus, it seems to be able to conclude that most of

the amino acid substitutions, no matter where they locate and what structure

change they may make, can cause TAP's highest tolerable temperature reduced

significantly. What's more, if the mutation occurring near or in the active

sites, it can also cause TAP's optimal reaction temperature reduced significantly

at the same time.

PMID: 15969072 [Indexed for MEDLINE]

3507. BMC Bioinformatics. 2003 Jun 23;4:25. Epub 2003 Jun 23.

AGRIS: Arabidopsis gene regulatory information server, an information resource of

Arabidopsis cis-regulatory elements and transcription factors.

Davuluri RV(1), Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold

E.

Author information:

(1)Human Cancer Genetics Program, Comprehensive Cancer Center, Dept of Molecular

Virology, Immunology and Medical Genetics, The Ohio State University, Columbus,

OH 43210, USA. Davuluri-1@medctr.osu.edu

BACKGROUND: The gene regulatory information is hardwired in the promoter regions

formed by cis-regulatory elements that bind specific transcription factors (TFs).

Hence, establishing the architecture of plant promoters is fundamental to

understanding gene expression. The determination of the regulatory circuits

controlled by each TF and the identification of the cis-regulatory sequences for

all genes have been identified as two of the goals of the Multinational

Coordinated Arabidopsis thaliana Functional Genomics Project by the Multinational

Arabidopsis Steering Committee (June 2002).

RESULTS: AGRIS is an information resource of Arabidopsis promoter sequences,

transcription factors and their target genes. AGRIS currently contains two

databases, AtTFDB (Arabidopsis thaliana transcription factor database) and

AtcisDB (Arabidopsis thaliana cis-regulatory database). AtTFDB contains

information on approximately 1,400 transcription factors identified through motif

searches and grouped into 34 families. AtTFDB links the sequence of the

transcription factors with available mutants and, when known, with the possible

genes they may regulate. AtcisDB consists of the 5' regulatory sequences of all

29,388 annotated genes with a description of the corresponding cis-regulatory

elements. Users can search the databases for (i) promoter sequences, (ii) a

transcription factor, (iii) a direct target genes for a specific transcription

factor, or (vi) a regulatory network that consists of transcription factors and

their target genes.

CONCLUSION: AGRIS provides the necessary software tools on Arabidopsis

transcription factors and their putative binding sites on all genes to initiate

the identification of transcriptional regulatory networks in the model

dicotyledoneous plant Arabidopsis thaliana. AGRIS can be accessed from

http://arabidopsis.med.ohio-state.edu.

DOI: 10.1186/1471-2105-4-25

PMCID: PMC166152

PMID: 12820902 [Indexed for MEDLINE]

3508. Bioinformatics. 2003 Jun 12;19(9):1153-4.

Coupled two-way clustering server.

Getz G(1), Domany E.

Author information:

(1)Department of Physics of Complex Systems, Weizmann Institute of Science,

Rehovot 76100, Israel. ctwc\_support@weizmann.ac.il

The CTWC server provides access to the software, CTWC1.00, that implements

Coupled Two Way Clustering (Getz et al., 2000), a method designed to mine gene

expression dataAVAILABILITY: Free, at http://ctwc.weizmann.ac.il.

SUPPLEMENTARY INFORMATION: The site has a link to an example which provides

figures and detailed explanations

PMID: 12801877 [Indexed for MEDLINE]

3509. J Digit Imaging. 2003 Jun;16(2):180-4. Epub 2003 Oct 2.

Implementing a MIRC query interface for a database driven teaching file.

Tellis WM(1), Andriole KP.

Author information:

(1)Laboratory for Radiological Informatics, University of California San

Francisco, 530 Parnassus Avenue, Room CL-158, San Francisco, CA 94143-0628, USA.

wyatt.tellis@radiology.ucsf.edu

This paper describes the authors' experience with integrating an existing

database-driven teaching file with the RSNA (Radiological Society of North

America) Medical Imaging Resource Center (MIRC). MIRC is the product of an

RSNA-sponsored initiative to enable medical institutions to share their

electronic medical content (images, text, and multimedia) by creating a

distributed repository accessible from the Internet. An existing database-driven

teaching file, developed by the authors and used extensively by the University of

California San Francisco (UCSF) Department of Radiology since 1998, was

retrofitted to include an interface for handling broadcast queries initiated by a

MIRC query service. These queries take place through the exchange of XML

documents via HTTP. After all the storage services have responded, the results

are collated by the query service and presented to the user. The teaching file

and MIRC interface were developed using the 4th Dimension Relational Database

Management System (RDBMS). The integration process primarily involved mapping the

"MIRCdocument" schema to the teaching file's schema, translating the actual MIRC

query into the internal query language of the database and extending the access

control mechanisms of the teaching file to allow public access. A working

implementation of the interface required only 3 days of development time, with

refinements taking place over several months. Interface development was greatly

aided by MIRC's use of well-established Internet standards. This project has

demonstrated the feasibility of implementing a MIRC interface on an existing

teaching file server.

DOI: 10.1007/s10278-003-1656-9

PMCID: PMC3046468

PMID: 14517722 [Indexed for MEDLINE]

3510. Bioinformatics. 2003 May 22;19(8):1037-8.

Visual representation of database search results: the RHIMS Plot.

Martin DM(1), Hill P, Barton GJ, Flavell AJ.

Author information:

(1)Post-Genomics and Molecular Interactions Centre, Wellcome Trust Biocentre,

University of Dundee, Dundee DD1 5EH, UK. d.m.a.martin@dundee.ac.uk

SUMMARY: An algorithm and software are described that provide a fast method to

produce a novel, function-oriented visualization of the results of a sequence

database search. Text mining of sequence annotations allows position specific

plots of potential functional similarity to be compared in a simple compact

representation.

AVAILABILITY: The application can be accessed via a web server at

http://www.compbio.dundee.ac.uk. The RHIMS software may be obtained by request to

the authors.

PMID: 12761069 [Indexed for MEDLINE]

3511. Bioinformatics. 2003 May 22;19(8):1015-8.

3D-Jury: a simple approach to improve protein structure predictions.

Ginalski K(1), Elofsson A, Fischer D, Rychlewski L.

Author information:

(1)BioInfoBank Institute, Limanowskiego 24A, 60-744 Poznan, Poland.

MOTIVATION: Consensus structure prediction methods (meta-predictors) have higher

accuracy than individual structure prediction algorithms (their components). The

goal for the development of the 3D-Jury system is to create a simple but powerful

procedure for generating meta-predictions using variable sets of models obtained

from diverse sources. The resulting protocol should help to improve the quality

of structural annotations of novel proteins.

RESULTS: The 3D-Jury system generates meta-predictions from sets of models

created using variable methods. It is not necessary to know prior characteristics

of the methods. The system is able to utilize immediately new components

(additional prediction providers). The accuracy of the system is comparable with

other well-tuned prediction servers. The algorithm resembles methods of selecting

models generated using ab initio folding simulations. It is simple and offers a

portable solution to improve the accuracy of other protein structure prediction

protocols.

AVAILABILITY: The 3D-Jury system is available via the Structure Prediction Meta

Server (http://BioInfo.PL/Meta/) to the academic community.

SUPPLEMENTARY INFORMATION: 3D-Jury is coupled to the continuous online server

evaluation program, LiveBench (http://BioInfo.PL/LiveBench/)

PMID: 12761065 [Indexed for MEDLINE]

3512. Bioinformatics. 2003 May 22;19(8):1009-14.

ProPred1: prediction of promiscuous MHC Class-I binding sites.

Singh H(1), Raghava GP.

Author information:

(1)Institute of Microbial technology, Chandigarh 160036, India.

SUMMARY: ProPred1 is an on-line web tool for the prediction of peptide binding to

MHC class-I alleles. This is a matrix-based method that allows the prediction of

MHC binding sites in an antigenic sequence for 47 MHC class-I alleles. The server

represents MHC binding regions within an antigenic sequence in user-friendly

formats. These formats assist user in the identification of promiscuous MHC

binders in an antigen sequence that can bind to large number of alleles. ProPred1

also allows the prediction of the standard proteasome and immunoproteasome

cleavage sites in an antigenic sequence. This server allows identification of MHC

binders, who have the cleavage site at the C terminus. The simultaneous

prediction of MHC binders and proteasome cleavage sites in an antigenic sequence

leads to the identification of potential T-cell epitopes.

AVAILABILITY: Server is available at http://www.imtech.res.in/raghava/propred1/.

Mirror site of this server is available at

http://bioinformatics.uams.edu/mirror/propred1/ SUPPLEMENTARY INFORMATION:

Matrices and document on server are available at

http://www.imtech.res.in/raghava/propred1/page2.html

PMID: 12761064 [Indexed for MEDLINE]

3513. Anat Rec B New Anat. 2003 May;272(1):91-7.

Using a modified standard microscope to generate virtual slides.

Romer DJ(1), Yearsley KH, Ayers LW.

Author information:

(1)Department of Pathology, The Ohio State University, Columbus 43210, USA.

romer.4@osu.edu

A standard microscope was reconfigured as a virtual slide generator by adding a

Prior Scientific H101 robotic stage with H29 controller and 0.1 microm linear

scales and a Hitachi HV-C20 3CCD camera. Media Cybernetics Image Pro Plus version

4 (IP4) software controlled stage movement in the X-, Y-, and Z-axis, whereas a

Media Cybernetics Pro-Series Capture Kit captured images at 640 x 480 pixels.

Stage calibration, scanning algorithms, storage requirements, and viewing modes

were standardized. IP4 was used to montage the captured images into a large

virtual slide image that was subsequently saved in TIF or JPEG format. Virtual

slides were viewed at the workstation using the IP4 viewer as well as Adobe

Photoshop and Kodak Imaging. MGI Zoom Server delivered the virtual slides to the

Internet, and MicroBrightField's Neuroinformatica viewing software provided a

browser-based virtual microscope interface together with labeling tools for

annotating virtual slides. The images were served from a Windows 2000 platform

with 2 GB RAM, 500 GB of disk storage, and a 1.0 GHz P4 processor. To conserve

disk space on the image server, TIF files were converted to the FlashPix (FPX)

file format using a compression ratio of 10:1. By using 4x, 10x, 20x, and 40x

objectives, very large gigapixel images of tissue whole-mounts and tissue arrays

with high quality and morphologic detail are now being generated for teaching,

publication, research, and morphometric analysis. Technical details and a

demonstration of our system can be found on the Web at

http://virtualmicroscope.osu.edu.

Copyright 2003 Wiley-Liss, Inc.

DOI: 10.1002/ar.b.10017

PMID: 12731075 [Indexed for MEDLINE]

3514. Bioinformatics. 2003 May 1;19(7):903-4.

IPPRED: server for proteins interactions inference.

Goffard N(1), Garcia V, Iragne F, Groppi A, de Daruvar A.

Author information:

(1)Centre de Bioinformatique Bordeaux, Université V. Segalen Bordeaux 2, 146 rue

Léo Saignat, France.

SUMMARY: IPPRED is a web based server to infer protein-protein interactions

through homology search between candidate proteins and those described as

interacting. This simple inference allows to propose or to validate potential

interactions.

AVAILABILITY: IPPRED is freely available at http://cbi.labri.fr/outils/ippred/.

PMID: 12724307 [Indexed for MEDLINE]

3515. Bioinformatics. 2003 May 1;19(7):899-900.

SCide: identification of stabilization centers in proteins.

Dosztányi Z(1), Magyar C, Tusnády G, Simon I.

Author information:

(1)Institute of Enzymology Biological Research Center, Hungarian Academy of

Sciences, PO Box 7, Budapest, H-1518, Hungary.

SUMMARY: SCide is a program to identify stabilization centers from known protein

structures. These are residues involved in cooperative long-range contacts, which

can be formed between various regions of a single polypeptide chain, or they can

belong to different peptides or polypeptides in a complex. The server takes a PDB

file as an input, and the result is presented in graphical or text format.

AVAILABILITY: SCide is available on the web at http://www.enzim.hu/scide. The

source code can be obtained from the authors on request.

PMID: 12724305 [Indexed for MEDLINE]

3516. J Biomol NMR. 2003 May;26(1):25-37.

PROSHIFT: protein chemical shift prediction using artificial neural networks.

Meiler J(1).

Author information:

(1)University of Washington, Department of Biochemistry, Box 357350, Seattle,

Washington 98195-7350, USA. jens@jens-meiler.de

The importance of protein chemical shift values for the determination of

three-dimensional protein structure has increased in recent years because of the

large databases of protein structures with assigned chemical shift data. These

databases have allowed the investigation of the quantitative relationship between

chemical shift values obtained by liquid state NMR spectroscopy and the

three-dimensional structure of proteins. A neural network was trained to predict

the (1)H, (13)C, and (15)N of proteins using their three-dimensional structure as

well as experimental conditions as input parameters. It achieves root mean square

deviations of 0.3 ppm for hydrogen, 1.3 ppm for carbon, and 2.6 ppm for nitrogen

chemical shifts. The model reflects important influences of the covalent

structure as well as of the conformation not only for backbone atoms (as, e.g.,

the chemical shift index) but also for side-chain nuclei. For protein models with

a RMSD smaller than 5 A a correlation of the RMSD and the r.m.s. deviation

between the predicted and the experimental chemical shift is obtained. Thus the

method has the potential to not only support the assignment process of proteins

but also help with the validation and the refinement of three-dimensional

structural proposals. It is freely available for academic users at the PROSHIFT

server: www.jens-meiler.de/proshift.html

PMID: 12766400 [Indexed for MEDLINE]

3517. Bioinformatics. 2003 Mar 1;19(4):543.

IS: a web-site for intron statistics.

Barta E(1), Kaján L, Pongor S.

Author information:

(1)Agricultural Biotechnology Center, 2100 Gödöllö, Hungary. barta@abc.hu

SUMMARY: A web server has been established for the statistical evaluation of

introns in various taxonomic groups and the comparison of taxonomic groups in

terms of intron type, length, base composition, etc. The options include the

graphic analysis of splice sites and a probability test for exon-shuffling within

the selected group.

AVAILABILITY: introns.abc.hu, http://www.icgeb.trieste.it/introns

PMID: 12611812 [Indexed for MEDLINE]

3518. Bioinformatics. 2003 Mar 1;19(4):506-12.

Detection of unrelated proteins in sequences multiple alignments by using

predicted secondary structures.

Errami M(1), Geourjon C, Deléage G.

Author information:

(1)Pôle de BioInformatique Lyonnais, Institut de Biologie et de Chimie des

Protéines, Centre National de la Recherche Scientifique, UMR 5086, 69367 Lyon

CEDEX 07, France.

MOTIVATION: Multiple sequence alignments are essential tools for establishing the

homology relations between proteins. Essential amino acids for the function

and/or the structure are generally conserved, thus providing key arguments to

help in protein characterization. However for distant proteins, it is more

difficult to establish, in a reliable way, the homology relations that may exist

between them. In this article, we show that secondary structure prediction is a

valuable way to validate protein families at low identity rate.

RESULTS: We show that the analysis of the secondary structures compatibility is a

reliable way to discard non-related proteins in low identity multiple alignment.

AVAILABILITY: This validation is possible through our NPS@ server

(http://npsa-pbil.ibcp.fr)

PMID: 12611806 [Indexed for MEDLINE]

3519. Bioinformatics. 2003 Mar 1;19(4):500-5.

MaxSubSeq: an algorithm for segment-length optimization. The case study of the

transmembrane spanning segments.

Fariselli P(1), Finelli M, Marchignoli D, Martelli PL, Rossi I, Casadio R.

Author information:

(1)Department of Biology/CIRB, University of Bologna, via Irnerio 42, 40126

Bologna, Italy. piero@biocomp.unibo.it

MOTIVATION: A problem in predicting the topography of transmembrane proteins is

the optimal localization of the transmembrane segments along the protein

sequences, provided that each residue is associated with a propensity of being or

not being included in the transmembrane protein region. From previous work it is

known that post-processing of propensity signals with suited algorithms can

greatly improve the quality and the accuracy of the predictions. In this paper we

describe a general dynamic programming-like algorithm (MaxSubSeq, Maximal

SubSequence) specifically designed to optimize the number and length of segments

with constrained length in a given protein sequence. Previous application of our

algorithm, has proved its effectiveness in the optimization task of both neural

network and hidden Markov models output, and in this paper we present the

detailed description of MaxSubSeq.

RESULTS: We describe the application of MaxSubSeq to the location of both helical

and beta strand transmembrane segments, optimizing the outputs derived with

different predictive algorithms. For all-alpha transmembrane proteins we use both

the standard Kyte-Doolittle (KD) hydropathy scale and the TMHMM predictor

(http://www.cbs.dtu.dk/). Using a set of 188 well characterized membrane

proteins, MaxSubSeq nearly doubles the correct location of transmembrane segments

as compared to the standard KD hydrophobicity plot, reaching 51% accuracy. If

MaxSubSeq is used to optimize the TMHMM method the accuracy increases from 68 to

72%. When used to regularize the prediction of beta transmembrane strands,

obtained using both a neural network and a HMM based predictors, MaxSubSeq

increases the accuracy per protein up to 72 and 73% respectively.

AVAILABILITY: The program is available upon request to the authors, or it is

accessible through our web server (http://gpcr.biocomp.unibo.it/predictors/)

PMID: 12611805 [Indexed for MEDLINE]

3520. Protein Sci. 2003 Mar;12(3):627-34.

Prediction of beta-turns in proteins from multiple alignment using neural

network.

Kaur H(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector 39A, Chandigarh, India.

A neural network-based method has been developed for the prediction of beta-turns

in proteins by using multiple sequence alignment. Two feed-forward

back-propagation networks with a single hidden layer are used where the

first-sequence structure network is trained with the multiple sequence alignment

in the form of PSI-BLAST-generated position-specific scoring matrices. The

initial predictions from the first network and PSIPRED-predicted secondary

structure are used as input to the second structure-structure network to refine

the predictions obtained from the first net. A significant improvement in

prediction accuracy has been achieved by using evolutionary information contained

in the multiple sequence alignment. The final network yields an overall

prediction accuracy of 75.5% when tested by sevenfold cross-validation on a set

of 426 nonhomologous protein chains. The corresponding Q(pred), Q(obs), and

Matthews correlation coefficient values are 49.8%, 72.3%, and 0.43, respectively,

and are the best among all the previously published beta-turn prediction methods.

The Web server BetaTPred2 (http://www.imtech.res.in/raghava/betatpred2/) has been

developed based on this approach.

DOI: 10.1110/ps.0228903

PMCID: PMC2312433

PMID: 12592033 [Indexed for MEDLINE]

3521. Proteins. 2003 Mar 1;50(4):600-8.

A neural network approach to evaluate fold recognition results.

Juan D(1), Graña O, Pazos F, Fariselli P, Casadio R, Valencia A.

Author information:

(1)Protein Design Group, National Center for Biotechnology, CNB-CSIC, Campus

Universidad Autónoma, Cantoblanco, Madrid, M-28049, Spain.

Fold recognition techniques assist the exploration of protein structures, and

web-based servers are part of the standard set of tools used in the analysis of

biochemical problems. Despite their success, current methods are only able to

predict the correct fold in a relatively small number of cases. We propose an

approach that improves the selection of correct folds from among the results of

two methods implemented as web servers (SAMT99 and 3DPSSM). Our approach is based

on the training of a system of neural networks with models generated by the

servers and a set of associated characteristics such as the quality of the

sequence-structure alignment, distribution of sequence features

(sequence-conserved positions and apolar residues), and compactness of the

resulting models. Our results show that it is possible to detect adequate folds

to model 80% of the sequences with a high level of confidence. The improvements

achieved by taking into account sequence characteristics open the door to future

improvements by directly including such factors in the step of model generation.

This approach has been implemented as an automatic system LIBELLULA, available as

a public web server at http://www.pdg.cnb.uam.es/servers/libellula.html.

Copyright 2003 Wiley-Liss, Inc.

DOI: 10.1002/prot.10322

PMID: 12577266 [Indexed for MEDLINE]

3522. Comput Methods Programs Biomed. 2003 Feb;70(2):99-105.

Use of correspondence discriminant analysis to predict the subcellular location

of bacterial proteins.

Perrière G(1), Thioulouse J.

Author information:

(1)Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS no 5558, Université

Claude Bernard-Lyon 1, 43, bd du 11 Novembre 1918, 69622 Villeurbanne Cedex,

France. perriere@biomserv.univ-lyon.fr

Correspondence discriminant analysis (CDA) is a multivariate statistical method

derived from discriminant analysis which can be used on contingency tables. We

have used CDA to separate Gram negative bacteria proteins according to their

subcellular location. The high resolution of the discrimination obtained makes

this method a good tool to predict subcellular location when this information is

not known. The main advantage of this technique is its simplicity. Indeed, by

computing two linear formulae on amino acid composition, it is possible to

classify a protein into one of the three classes of subcellular location we have

defined. The CDA itself can be computed with the ADE-4 software package that can

be downloaded, as well as the data set used in this study, from the Pôle

Bio-Informatique Lyonnais (PBIL) server at http://pbil.univ-lyon1.fr.

PMID: 12507786 [Indexed for MEDLINE]

3523. Bioinformatics. 2003 Jan 22;19(2):219-27.

Gene finding with a hidden Markov model of genome structure and evolution.

Pedersen JS(1), Hein J.

Author information:

(1)Bioinformatics Research Center, Department of Genetics and Ecology, The

Institute of Biological Sciences, University of Aarhus, Building 550, Ny

Munkegade, 8000 Aarhus C, Denmark. jsp@daimi.au.dk

MOTIVATION: A growing number of genomes are sequenced. The differences in

evolutionary pattern between functional regions can thus be observed genome-wide

in a whole set of organisms. The diverse evolutionary pattern of different

functional regions can be exploited in the process of genomic annotation. The

modelling of evolution by the existing comparative gene finders leaves room for

improvement.

RESULTS: A probabilistic model of both genome structure and evolution is

designed. This type of model is called an Evolutionary Hidden Markov Model

(EHMM), being composed of an HMM and a set of region-specific evolutionary models

based on a phylogenetic tree. All parameters can be estimated by maximum

likelihood, including the phylogenetic tree. It can handle any number of aligned

genomes, using their phylogenetic tree to model the evolutionary correlations.

The time complexity of all algorithms used for handling the model are linear in

alignment length and genome number. The model is applied to the problem of gene

finding. The benefit of modelling sequence evolution is demonstrated both in a

range of simulations and on a set of orthologous human/mouse gene pairs.

AVAILABILITY: Free availability over the Internet on www server:

http://www.birc.dk/Software/evogene.

PMID: 12538242 [Indexed for MEDLINE]

3524. Bioinformatics. 2003;19 Suppl 1:i252-4.

STRUCLA: a WWW meta-server for protein structure comparison and evolutionary

classification.

Sasin JM(1), Kurowski MA, Bujnicki JM.

Author information:

(1)Bioinformatics Laboratory, International Institute of Molecular and Cell

Biology, Trojdena 4, 02-109 Warsaw, Poland.

MOTIVATION: Evolutionary relationships of proteins have long been derived from

the alignment of protein sequences. But from the view of function, most

restraints of evolutionary divergence operate at the level of tertiary structure.

It has been demonstrated that quantitative measures of dissimilarity in families

of structurally similar proteins can be applied to the construction of trees from

a comparison of their three-dimensional structures. However, no convenient tool

is publicly available to carry out such analyses.

RESULTS: We developed STRUCLA (STRUcture CLAssification), a WWW tool for

generation of trees based on evolutionary distances inferred from protein

structures according to various methods. The server takes as an input a list of

PDB files or the initial alignment of protein coordinates provided by the user

(for instance exported from SWISS PDB VIEWER). The user specifies the distance

cutoff and selects the distance measures. The server returns series of unrooted

trees in the NEXUS format and corresponding distance matrices, as well as a

consensus tree. The results can be used as an alternative and a complement to a

fixed hierarchy of current protein structure databases. It can complement

sequence-based phylogenetic analysis in the 'twilight zone of homology', where

amino acid sequences are too diverged to provide reliable relationships.

PMID: 12855467 [Indexed for MEDLINE]

3525. Bioinformatics. 2003;19 Suppl 1:i250-1.

A tool to assist the study of specific features at protein binding sites.

Santander V(1), Portales MA, Melo F.

Author information:

(1)Pontificia Universidad Católica de Chile, Departmento Genética Molecular y

Microbiología, Facultad de Ciencias Biológicas, Alameda 340, Santiago, Chile.

The Protein Data Bank contains a large amount of proteins that have been solved

with small ligands bound to them. This constitutes a rich source of information

for the study of the specific requirements of protein sites to bind small

molecules with a favorable free energy. The specific atomic composition and

three-dimensional geometric restraints of protein binding sites for different

ligands could be easily obtained from there. The development of accurate binding

site descriptors in proteins constitutes a valuable tool to assist in the

large-scale prediction and annotation of protein function in whole genomes. In

this work, an integrated database containing some processed and calculated

protein/ligand information is described. It is expected that this database will

constitute a useful tool for people working in the prediction of protein function

from its structure. The database is accessible from the Internet through a web

server located at: http://protein.bio.puc.cl

PMID: 12855466 [Indexed for MEDLINE]

3526. Bioinformatics. 2003 Jan;19(1):167-8.

Detection of hydrogen-bond signature patterns in protein families.

Prasad T(1), Prathima MN, Chandra N.

Author information:

(1)Bioinformatics Centre, Supercomputer Education and Research Centre, Indian

Institute of Science, Bangalore 560012, India.

SUMMARY: We have developed a WWW server, HBPRINT, for the identification of

hydrogen-bond signature patterns in protein families from their structures. The

server calculates (a) common hydrogen bonds between two structures (b) a

hydrogen-bond fingerprint in a set of structural neighbours and (c) details of

conserved hydrogen bonds. The server also enables the visualization of the

hydrogen bond network comprising the signature pattern.

AVAILABILITY: HBPRINT and a tutorial are available from

http://144.16.93.115/hb\_page/index.html.

PMID: 12499314 [Indexed for MEDLINE]

3527. Bioinformatics. 2003 Jan;19(1):163-4.

ConSurf: identification of functional regions in proteins by surface-mapping of

phylogenetic information.

Glaser F(1), Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N.

Author information:

(1)The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv

69978, Israel. fabian@ashtoret.tau.ac.il

We recently developed algorithmic tools for the identification of functionally

important regions in proteins of known three dimensional structure by estimating

the degree of conservation of the amino-acid sites among their close sequence

homologues. Projecting the conservation grades onto the molecular surface of

these proteins reveals patches of highly conserved (or occasionally highly

variable) residues that are often of important biological function. We present a

new web server, ConSurf, which automates these algorithmic tools. ConSurf may be

used for high-throughput characterization of functional regions in

proteins.AVAILABILITY: The ConSurf web server is available

at:http://consurf.tau.ac.il.

SUPPLEMENTARY INFORMATION: A set of examples is available at

http://consurf.tau.ac.il under 'GALLERY'.

PMID: 12499312 [Indexed for MEDLINE]

3528. Bioinformatics. 2003 Jan;19(1):87-9.

MADGE: scalable distributed data management software for cDNA microarrays.

McIndoe RA(1), Lanzen A, Hurtz K.

Author information:

(1)Department of Pathology, Immunology and Laboratory Medicine, University of

Florida, Gainesville 32610, USA. rmcindoe@mail.mcg.edu

MOTIVATION: The human genome project and the development of new high-throughput

technologies have created unparalleled opportunities to study the mechanism of

diseases, monitor the disease progression and evaluate effective therapies. Gene

expression profiling is a critical tool to accomplish these goals. The use of

nucleic acid microarrays to assess the gene expression of thousands of genes

simultaneously has seen phenomenal growth over the past five years. Although

commercial sources of microarrays exist, investigators wanting more flexibility

in the genes represented on the array will turn to in-house production. The

creation and use of cDNA microarrays is a complicated process that generates an

enormous amount of information. Effective data management of this information is

essential to efficiently access, analyze, troubleshoot and evaluate the

microarray experiments.

RESULTS: We have developed a distributable software package designed to track and

store the various pieces of data generated by a cDNA microarray facility. This

includes the clone collection storage data, annotation data, workflow queues,

microarray data, data repositories, sample submission information, and

project/investigator information. This application was designed using a 3-tier

client server model. The data access layer (1st tier) contains the relational

database system tuned to support a large number of transactions. The data

services layer (2nd tier) is a distributed COM server with full database

transaction support. The application layer (3rd tier) is an internet based user

interface that contains both client and server side code for dynamic interactions

with the user.

AVAILABILITY: This software is freely available to academic institutions and

non-profit organizations at http://www.genomics.mcg.edu/niddkbtc.

PMID: 12499297 [Indexed for MEDLINE]

3529. Bioinformatics. 2003 Jan;19(1):79-86.

Mining gene expression databases for association rules.

Creighton C(1), Hanash S.

Author information:

(1)Bioinformatics Program Pediatrics and Communicable Diseases, University of

Michigan, Ann Arbor 48109, USA. ccreight@umich.edu

MOTIVATION: Global gene expression profiling, both at the transcript level and at

the protein level, can be a valuable tool in the understanding of genes,

biological networks, and cellular states. As larger and larger gene expression

data sets become available, data mining techniques can be applied to identify

patterns of interest in the data. Association rules, used widely in the area of

market basket analysis, can be applied to the analysis of expression data as

well. Association rules can reveal biologically relevant associations between

different genes or between environmental effects and gene expression. An

association rule has the form LHS --> RHS, where LHS and RHS are disjoint sets of

items, the RHS set being likely to occur whenever the LHS set occurs. Items in

gene expression data can include genes that are highly expressed or repressed, as

well as relevant facts describing the cellular environment of the genes (e.g. the

diagnosis of a tumor sample from which a profile was obtained).

RESULTS: We demonstrate an algorithm for efficiently mining association rules

from gene expression data, using the data set from Hughes et al. (2000, Cell,

102, 109-126) of 300 expression profiles for yeast. Using the algorithm, we find

numerous rules in the data. A cursory analysis of some of these rules reveals

numerous associations between certain genes, many of which make sense

biologically, others suggesting new hypotheses that may warrant further

investigation. In a data set derived from the yeast data set, but with the

expression values for each transcript randomly shifted with respect to the

experiments, no rules were found, indicating that most all of the rules mined

from the actual data set are not likely to have occurred by chance.

AVAILABILITY: An implementation of the algorithm using Microsoft SQL Server with

Access 2000 is available at

http://dot.ped.med.umich.edu:2000/pub/assoc\_rules/assoc\_rules.zip. Our results

from mining the yeast data set are available at

http://dot.ped.med.umich.edu:2000/pub/assoc\_rules/yeast\_results.zip.

PMID: 12499296 [Indexed for MEDLINE]

3530. Comp Funct Genomics. 2003;4(4):420-3. doi: 10.1002/cfg.309.

In silico identification of functional protein interfaces.

Bell RE(1), Ben-Tal N.

Author information:

(1)Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel

Aviv University, Ramat Aviv 69978, Israel.

Proteins perform many of their biological roles through protein-protein,

protein-DNA or protein-ligand interfaces. The identification of the amino acids

comprising these interfaces often enhances our understanding of the biological

function of the proteins. Many methods for the detection of functional interfaces

have been developed, and large-scale analyses have provided assessments of their

accuracy. Among them are those that consider the size of the protein interface,

its amino acid composition and its physicochemical and geometrical properties.

Other methods to this effect use statistical potential functions of pairwise

interactions, and evolutionary information. The rationale of the evolutionary

approach is that functional and structural constraints impose selective pressure;

hence, biologically important interfaces often evolve at a slower pace than do

other external regions of the protein. Recently, an algorithm, Rate4Site, and a

web-server, ConSurf (http://consurf.tau.ac.il/), for the identification of

functional interfaces based on the evolutionary relations among homologous

proteins as reflected in phylogenetic trees, were developed in our laboratory.

The explicit use of the tree topology and branch lengths makes the method

remarkably accurate and sensitive. Here we demonstrate its potency in the

identification of the functional interfaces of a hypothetical protein, the

structure of which was determined as part of the international structural

genomics effort. Finally, we propose to combine complementary procedures, in

order to enhance the overall performance of methods for the identification of

functional interfaces in proteins.

DOI: 10.1002/cfg.309

PMCID: PMC2447364

PMID: 18629079

3531. In Silico Biol. 2003;3(4):405-9.

secureBLAST.

Wiezer A(1), Merkl R.

Author information:

(1)Goettingen Genomics Laboratory, Institut fuer Mikrobiologie and Genetik,

Grisebachstrasse 8, D-37077 Goettingen, Germany.

secureBLAST supplements NCBI wwwblast with features necessary to control in an

easy manageable way usage of BLAST data sets and their update. The concept we

implemented allows to offer on a single BLAST server several data sets with

individually configurable access rights. Security is provided by user

authentication and encryption of the http traffic via SSL. By using secureBLAST,

the administration of users and databases can be done via a web interface.

Therefore, secureBLAST is valuable for institutions that have to restrict access

to their datasets or just want to administer BLAST servers via a web interface.

PMID: 12954083 [Indexed for MEDLINE]

3532. Int J Med Inform. 2003 Jan;69(1):57-62.

International distance-learning outreach: the APEC EINet experience.

Kimball AM(1), Shih L, Brown J, Harris TG, Pautler N, Jamieson RW, Bolles J,

Horwitch C.

Author information:

(1)Department of Health Services, University of Washington, Seattle, WA, USA.

akimball@u.washington.edu

BACKGROUND: The Emerging Infections Network is a mature electronic network that

links Public Health professionals in the Asia Pacific through regular e-mail

bulletins and an extensive Web site (http://www.apec.org/infectious). Emerging

infections is a new area of study; learning materials help foster education. Our

objective is to quantify the response of the network to the introduction of

distance-learning materials on the Web site.

METHODS: Distance-learning materials, developed by the University of Washington

School of Public Health, were field tested and launched on the site. Publicity

was carried out prior to the launch of the materials. Access was tracked

prospectively using server counts of page downloads.

RESULTS: Web access increased substantially during the month after the materials

were launched, especially among Asia based computers. The effect was isolated to

the distance-learning pages, and not general to the site.

CONCLUSIONS: This Web site appears to be responsive to the advertisement and to

the materials. Prospective Web-site monitoring proved useful.

Copyright 2002 Elsevier Science Ireland Ltd.

PMID: 12485704 [Indexed for MEDLINE]

3533. J Struct Funct Genomics. 2003;4(2-3):121-7.

Secure web book to store structural genomics research data.

Manjasetty BA(1), Höppner K, Mueller U, Heinemann U.

Author information:

(1)Protein Structure Factory, c/o BESSY GmbH, Albert-Einstein Str.15, 12489,

Berlin, Germany.

Recently established collaborative structural genomics programs aim at

significantly accelerating the crystal structure analysis of proteins. These

large-scale projects require efficient data management systems to ensure seamless

collaboration between different groups of scientists working towards the same

goal. Within the Berlin-based Protein Structure Factory, the synchrotron X-ray

data collection and the subsequent crystal structure analysis tasks are located

at BESSY, a third-generation synchrotron source. To organize file-based

communication and data transfer at the BESSY site of the Protein Structure

Factory, we have developed the web-based BCLIMS, the BESSY Crystallography

Laboratory Information Management System. BCLIMS is a relational data management

system which is powered by MySQL as the database engine and Apache HTTP as the

web server. The database interface routines are written in Python programing

language. The software is freely available to academic users. Here we describe

the storage, retrieval and manipulation of laboratory information, mainly

pertaining to the synchrotron X-ray diffraction experiments and the subsequent

protein structure analysis, using BCLIMS.

PMID: 14649296 [Indexed for MEDLINE]

3534. Nucleic Acids Res. 2003 Jan 1;31(1):478-82.

MolMovDB: analysis and visualization of conformational change and structural

flexibility.

Echols N(1), Milburn D, Gerstein M.

Author information:

(1)Department of Molecular Biophysics and Biochemistry, Yale University, 266

Whitney Avenue, New Haven, CT 06520, USA.

The Database of Macromolecular Movements (http://MolMovDB.org) is a collection of

data and software pertaining to flexibility in protein and RNA structures. The

database is organized into two parts. Firstly, a collection of 'morphs' of solved

structures representing different states of a molecule provides quantitative data

for flexibility and a number of graphical representations. Secondly, a

classification of known motions according to type of conformational change (e.g.

'hinged domain' or 'allosteric') incorporates textual annotation and information

from the literature relating to the motion, linking together many of the morphs.

A variety of subsets of the morphs are being developed for use in statistical

analyses. In particular, for each subset it is possible to derive distributions

of various motional quantities (e.g. maximum rotation) that can be used to place

a specific motion in context as being typical or atypical for a given population.

Over the past year, the database has been greatly expanded and enhanced to

incorporate new structures and to improve the quality of data. The 'morph

server', which enables users of the database to add new morphs either from their

own research or the PDB, has also been enhanced to handle nucleic acid structures

and multi-chain complexes.

PMCID: PMC165551

PMID: 12520056 [Indexed for MEDLINE]

3535. Nucleic Acids Res. 2003 Jan 1;31(1):363-4.

SRPDB: Signal Recognition Particle Database.

Rosenblad MA(1), Gorodkin J, Knudsen B, Zwieb C, Samuelsson T.

Author information:

(1)Department of Medical Biochemistry, University of Göteborg, Box 440, SE-405 30

Göteborg, Sweden.

The Signal Recognition Particle Database (SRPDB) at

http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html and

http://bio.lundberg.gu.se/dbs/SRPDB/SRPDB.html assists in the better

understanding of the structure and function of the signal recognition particle

(SRP), a ribonucleoprotein complex that recognizes signal sequences as they

emerge from the ribosome. SRPDB provides alphabetically and phylogenetically

ordered lists of SRP RNA and SRP protein sequences. The SRP RNA alignment

emphasizes base pairs supported by comparative sequence analysis to derive

accurate SRP RNA secondary structures for each species. This release includes a

total of 181 SRP RNA sequences, 7 protein SRP9, 11 SRP14, 31 SRP19, 113 SRP54

(Ffh), 9 SRP68 and 12 SRP72 sequences. There are 44 new sequences of the SRP

receptor alpha subunit and its FtsY homolog (a total of 99 entries). Additional

data are provided for polypeptides with established or potential roles in

SRP-mediated protein targeting, such as the beta subunit of SRP receptor, Flhf,

Hbsu and cpSRP43. Also available are motifs for the identification of new SRP RNA

sequences, 2D representations, three-dimensional models in PDB format, and links

to the high-resolution structures of several SRP components. New to this version

of SRPDB is the introduction of a relational database system and a SRP RNA

prediction server (SRP-Scan) which allows the identification of SRP RNAs within

genome sequences and also generates secondary structure diagrams.

PMCID: PMC165554

PMID: 12520023 [Indexed for MEDLINE]

3536. Nucleic Acids Res. 2003 Jan 1;31(1):359-62.

SDAP: database and computational tools for allergenic proteins.

Ivanciuc O(1), Schein CH, Braun W.

Author information:

(1)Sealy Center for Structural Biology, Department of Human Biological Chemistry

and Genetics, University of Texas Medical Branch, 310 University Boulevard,

Galveston, TX 77555-1157, USA.

SDAP (Structural Database of Allergenic Proteins) is a web server that provides

rapid, cross-referenced access to the sequences, structures and IgE epitopes of

allergenic proteins. The SDAP core is a series of CGI scripts that process the

user queries, interrogate the database, perform various computations related to

protein allergenic determinants and prepare the output HTML pages. The database

component of SDAP contains information about the allergen name, source, sequence,

structure, IgE epitopes and literature references and easy links to the major

protein (PDB, SWISS-PROT/TrEMBL, PIR-ALN, NCBI Taxonomy Browser) and literature

(PubMed, MEDLINE) on-line servers. The computational component in SDAP uses an

original algorithm based on conserved properties of amino acid side chains to

identify regions of known allergens similar to user-supplied peptides or selected

from the SDAP database of IgE epitopes. This and other bioinformatics tools can

be used to rapidly determine potential cross-reactivities between allergens and

to screen novel proteins for the presence of IgE epitopes they may share with

known allergens. SDAP is available via the World Wide Web at

http://fermi.utmb.edu/SDAP/.

PMCID: PMC165457

PMID: 12520022 [Indexed for MEDLINE]

3537. Nucleic Acids Res. 2003 Jan 1;31(1):353-8.

RTKdb: database of Receptor Tyrosine Kinase.

Grassot J(1), Mouchiroud G, Perrière G.

Author information:

(1)Centre de Génétique Moléculaire et Cellulaire, UMR CNRS 5534, Université

Claude Bernard-Lyon 1, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex,

France. grassot@biomserv.univ-lyon1.fr

Receptor Tyrosine Kinases (RTK) are transmembrane receptors specifically found in

metazoans. They represent an excellent model for studying evolution of cellular

processes in metazoans because they encompass large families of modular proteins

and belong to a major family of contingency generating molecules in eukaryotic

cells: the protein kinases. Because tyrosine kinases have been under close

scrutiny for many years in various species, they are associated with a wealth of

information, mainly in mammals. Presently, most categories of RTK were identified

in mammals, but in a near future other model species will be sequenced, and will

bring us RTKs from other metazoan clades. Thus, collecting RTK sequences would

provide a good starting point as a new model for comparative and evolutionary

studies applying to multigene families. In this context, we are developing the

Receptor Tyrosine Kinase database (RTKdb), which is the only database on tyrosine

kinase receptors presently available. In this database, protein sequences from

eight model metazoan species are organized under the format previously used for

the HOVERGEN, HOBACGEN and NUREBASE systems. RTKdb can be accessed through the

PBIL (Pôle Bioinformatique Lyonnais) World Wide Web server at

http://pbil.univ-lyon1.fr/RTKdb/, or through the FamFetch graphical user

interface available at the same address.

PMCID: PMC165483

PMID: 12520021 [Indexed for MEDLINE]

3538. Nucleic Acids Res. 2003 Jan 1;31(1):207-11.

The PEDANT genome database.

Frishman D(1), Mokrejs M, Kosykh D, Kastenmüller G, Kolesov G, Zubrzycki I,

Gruber C, Geier B, Kaps A, Albermann K, Volz A, Wagner C, Fellenberg M, Heumann

K, Mewes HW.

Author information:

(1)Institute for Bioinformatics, GSF - National Research Center for Environment

and Health, Ingolstädter Landstrasse 1, 85764 Neueherberg, Germany.

d.frishman@gsf.de

The PEDANT genome database (http://pedant.gsf.de) provides exhaustive automatic

analysis of genomic sequences by a large variety of established bioinformatics

tools through a comprehensive Web-based user interface. One hundred and seventy

seven completely sequenced and unfinished genomes have been processed so far,

including large eukaryotic genomes (mouse, human) published recently. In this

contribution, we describe the current status of the PEDANT database and novel

analytical features added to the PEDANT server in 2002. Those include: (i)

integration with the BioRS data retrieval system which allows fast text queries,

(ii) pre-computed sequence clusters in each complete genome, (iii) a

comprehensive set of tools for genome comparison, including genome comparison

tables and protein function prediction based on genomic context, and (iv)

computation and visualization of protein-protein interaction (PPI) networks based

on experimental data. The availability of functional and structural predictions

for 650 000 genomic proteins in well organized form makes PEDANT a useful

resource for both functional and structural genomics.

PMCID: PMC165452

PMID: 12519983 [Indexed for MEDLINE]

3539. Nucleic Acids Res. 2003 Jan 1;31(1):106-8.

MICdb: database of prokaryotic microsatellites.

Sreenu VB(1), Alevoor V, Nagaraju J, Nagarajaram HA.

Author information:

(1)Laboratory of Computational Biology and Bioinformaitcs Facility, ECIL Road,

Nacharam, Hyderabad 500 076, India.

The MICdb (Microsatellites Database) (http://www.cdfd.org.in/micas) is a

comprehensive relational database of non-redundant microsatellites extracted from

fully sequenced prokaryotic genomes. The current version (1.0) of the database

has been compiled from 83 genomes belonging to different phylogenetic groups.

This database has been linked to MICAS, the web-based Microstatellite Analysis

Server. MICAS provides a user-friendly front-end to systematically extract data

on microsatellite tracts from genomes. The database contains the following

information pertaining to the microsatellites: the regions (coding/non-coding, if

coding, their GenBank annotations) containing microsatellite tracts; the

frequencies of their occurrences, the size and the number of repeating motifs;

and the sequences of the tracts. MICAS also provides an interface to Autoprimer,

a primer design program to automatically design primers for selected

microsatellite loci.

PMCID: PMC165449

PMID: 12519959 [Indexed for MEDLINE]

3540. Pol J Pathol. 2003;54(3):223-6.

Electronic patient record and archive of records in Cardio.net system for

telecardiology.

Sierdziński J(1), Karpiński G.

Author information:

(1)Department of Medical Informatics, Medical University, Warszawa.

In modern medicine the well structured patient data set, fast access to it and

reporting capability become an important question. With the dynamic development

of information technology (IT) such question is solved via building electronic

patient record (EPR) archives. We then obtain fast access to patient data,

diagnostic and treatment protocols etc. It results in more efficient, better and

cheaper treatment. The aim of the work was to design a uniform Electronic Patient

Record, implemented in cardio.net system for telecardiology allowing the

co-operation among regional hospitals and reference centers. It includes

questionnaires for demographic data and questionnaires supporting doctor's work

(initial diagnosis, final diagnosis, history and physical, ECG at the discharge,

applied treatment, additional tests, drugs, daily and periodical reports). The

browser is implemented in EPR archive to facilitate data retrieval. Several tools

for creating EPR and EPR archive were used such as: XML, PHP, Java Script and

MySQL. The separate question is the security of data on WWW server. The security

is ensured via Security Socket Layer (SSL) protocols and other tools. EPR in

Cardio.net system is a module enabling the co-work of many physicians and the

communication among different medical centers.

PMID: 14703293 [Indexed for MEDLINE]

3541. Proc IEEE Comput Soc Bioinform Conf. 2003;2:180-9.

Statistical and visual morph movie analysis of crystallographic mutant selection

bias in protein mutation resource data.

Krebs WG(1), Bourne PE.

Author information:

(1)Department of Pharmacology, University of California at San Diego, La Jolla,

92093-0505, USA. wkrebs@sdsc.edu

The relationship between protein mutations and conformational change can

potentially decipher the language relating sequence to structure. Elsewhere, we

presented the Protein Mutant Resource (PMR), an online tool that systematically

identified related mutants in the Protein DataBank (PDB), inferred mutant Gene

Ontology classifications using data-mining, and allowed intuitive exploration of

relationships between mutant structures. Here, we perform a comprehensive

statistical analysis of PMR mutants. Although the PMR contains spectacular

conformational changes, generally there is a counter-intuitive inverse

relationship between conformational change and the number of mutations. That is,

PDB mutations contrast naturally evolved mutations. We compare the frequencies of

mutations in the PMR/PDB datasets against the PAM250 natural mutation frequencies

to confirm this. We make available morph movies from PMR structure pairs,

allowing visual analysis of conformational change and the ability to distinguish

visually between conformational change due to motions (e.g., ligand binding)and

mutations. The PMR is at http://pmr.sdsc.edu.

PMID: 16452792 [Indexed for MEDLINE]

3542. Proteins. 2003;53 Suppl 6:548-60.

CAFASP3 in the spotlight of EVA.

Eyrich VA(1), Przybylski D, Koh IY, Grana O, Pazos F, Valencia A, Rost B.

Author information:

(1)CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia

University, New York, New York 10032, USA.

We have analysed fold recognition, secondary structure and contact prediction

servers from CAFASP3. This assessment was carried out in the framework of the

fully automated, web-based evaluation server EVA. Detailed results are available

at http://cubic.bioc.columbia.edu/eva/cafasp3/. We observed that the

sequence-unique targets from CAFASP3/CASP5 were not fully representative for

evaluating performance. For all three categories, we showed how careless ranking

might be misleading. We compared methods from all categories to experts in

secondary structure and contact prediction and homology modellers to fold

recognisers. While the secondary structure experts clearly outperformed all

others, the contact experts appeared to outperform only novel fold methods.

Automatic evaluation servers are good at getting statistics right and at using

these to discard misleading ranking schemes. We challenge that to let machines

rule where they are best might be the best way for the community to enjoy the

tremendous benefit of CASP as a unique opportunity for brainstorming.

Copyright 2003 Wiley-Liss, Inc.

DOI: 10.1002/prot.10534

PMID: 14579345 [Indexed for MEDLINE]

3543. Proteins. 2003;53 Suppl 6:369-79.

A "FRankenstein's monster" approach to comparative modeling: merging the finest

fragments of Fold-Recognition models and iterative model refinement aided by 3D

structure evaluation.

Kosinski J(1), Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM.

Author information:

(1)Bioinformatics Laboratory, International Institute of Molecular and Cell

Biology, Trojdena 4, 02-109 Warsaw, Poland.

We applied a new multi-step protocol to predict the structures of all targets

during CASP5, regardless of their potential category. 1) We used diverse

fold-recognition (FR) methods to generate initial target-template alignments,

which were converted into preliminary full-atom models by comparative modeling.

All preliminary models were evaluated (scored) by VERIFY3D to identify well- and

poorly-folded fragments. 2) Preliminary models with similar 3D folds were

superimposed, poorly-scoring regions were deleted and the "average model"

structure was created by merging the remaining segments. All template structures

reported by FR were superimposed and a composite multiple-structure template was

created from the most conserved fragments. 3). The average model was superimposed

onto the composite template and the structure-based target-template alignment was

inferred. This alignment was used to build a new (intermediate) comparative model

of the target, again scored with VERIFY3D. 4) For all poorly scoring regions

series of alternative alignments were generated by progressively shifting the

"unfit" sequence fragment in either direction. Here, we considered additional

information, such as secondary structure, placement of insertions and deletions

in loops, conservation of putative catalytic residues, and the necessity to

obtain a compact, well-folded structure. For all alternative alignments, new

models were built and evaluated. 5) All models were superimposed and the

"FRankenstein's monster" (FR, fold recognition) model was built from best-scoring

segments. The final model was obtained after limited energy minimization to

remove steric clashes between sidechains from different fragments. The novelty of

this approach is in the focus on "vertical" recombination of structure fragments,

typical for the ab initio field, rather than "horizontal" sequence alignment

typical for comparative modeling. We tested the usefulness of the "FRankenstein"

approach for non-expert predictors: only the leader of our team had considerable

experience in protein modeling - he registered as a separate group (020) and

submitted models built only by himself. At the onset of CASP5, the other five

members of the team (students) had very little or no experience with modeling.

They followed the same protocol in a deliberately naïve way. In the fourth step

they used solely the VERIFY3D criterion to compare their models and the leader's

model (the latter regarded only as one of the many alternatives) and generated

the hybrid or selected only one model for submission (group 517). In order to

compare our protocol with the traditional "one target-one template-one alignment"

approach, we submitted (as a separate group 242) models selected from those

automatically generated by all CAFASP servers (i.e. obtained without any human

intervention). Here, we compare the results obtained by the three "groups",

describe successes and failures of the "FRankenstein" approach and discuss future

developments of comparative modeling. The automatic version of our multi-step

protocol is being developed as a meta-server; the prototype is freely available

at http://genesilico.pl/meta/.

Copyright 2003 Wiley-Liss, Inc.

DOI: 10.1002/prot.10545

PMID: 14579325 [Indexed for MEDLINE]

3544. Water Sci Technol. 2003;47(5):31-7.

Automated biofilm morphology quantification from confocal laser scanning

microscopy imaging.

Xavier JB(1), White DC, Almeida JS.

Author information:

(1)ITQB/UNL, R Qta Grande 6, 2780 Oeiras, Portugal.

In spite of the immediate visual appeal of confocal laser scanning microscopy

images, the extraction of accurate reconstitutions of biofilm morphology requires

a lengthy and computational intensive succession of processing steps. However,

once performed, it provides ample reward by enabling the quantitative study of

biofilm structure. A software suite of image processing tools for full automation

of biofilm morphology quantification was developed by integrating preprocessing,

segmentation and morphology quantification operations. This software toolbox was

implemented in a web server and a user friendly interface was developed to

facilitate image submission, storage and sharing, its access being unrestricted

for scientific applications. The image bioinformatics tool which results from the

integration of the processing operations can be accessed at

http://www.itqb.unl.pt:111/clsmip/. Its use is described in this paper and is

illustrated with an example of processing of experimental data describing the

growth of a mixed species dentrifying biofilm.

PMID: 12701903 [Indexed for MEDLINE]

3545. Acta Crystallogr B. 2002 Dec;58(Pt 6):921-33. Epub 2002 Nov 28.

Prediction of new displacive ferroelectrics through systematic pseudosymmetry

search. Results for materials with Pba2 and Pmc21 symmetry.

Kroumova E(1), Aroyo MI, Perez-Mato JM.

Author information:

(1)Departamento de Fisica de la Materia Condensada, Universidad del Pais Vasco,

Apdo 644, 48080 Bilbao, Spain.

Polar structures with pseudosymmetry related to a hypothetical non-polar

configuration are considered as good candidates for ferroelectrics. Recently, a

procedure has been developed for a systematic pseudosymmetry search among

structures with a given space-group symmetry. The aim of this paper is the

extension of the pseudosymmetry procedure to the case of structures with polar

symmetry and its application in the search for new ferroelectrics. The results

obtained by the generalized pseudosymmetry search among the compounds with

symmetries Pba2 and Pmc2(1) listed in the Inorganic Crystal Structure Database

are discussed. The calculations have been performed by the program PSEUDO, which

forms part of the Bilbao Crystallographic Server (http://www.cryst.ehu.es). In

addition, an empirical relation between the atomic displacements necessary to

reach the non-polar structure and the transition temperature is proposed and

compared with the Abrahams-Kurtz-Jamieson relation.

PMID: 12456970

3546. Genomics. 2002 Dec;80(6):681-90.

PipTools: a computational toolkit to annotate and analyze pairwise comparisons of

genomic sequences.

Elnitski L(1), Riemer C, Petrykowska H, Florea L, Schwartz S, Miller W, Hardison

R.

Author information:

(1)Department of Computer Science and Engineering, The Pennsylvania State

University, University Park, PA 16802, USA. elnitski@bio.cse.psu.edu

Sequence conservation between species is useful both for locating coding regions

of genes and for identifying functional noncoding segments. Hence interspecies

alignment of genomic sequences is an important computational technique. However,

its utility is limited without extensive annotation. We describe a suite of

software tools, PipTools, and related programs that facilitate the annotation of

genes and putative regulatory elements in pairwise alignments. The alignment

server PipMaker uses the output of these tools to display detailed information

needed to interpret alignments. These programs are provided in a portable format

for use on common desktop computers and both the toolkit and the PipMaker server

can be found at our Web site (http://bio.cse.psu.edu/). We illustrate the utility

of the toolkit using annotation of a pairwise comparison of the mouse MHC class

II and class III regions with orthologous human sequences and subsequently

identify conserved, noncoding sequences that are DNase I hypersensitive sites in

chromatin of mouse cells.

PMID: 12504859 [Indexed for MEDLINE]

3547. Hear Res. 2002 Dec;174(1-2):86-92.

Radiation hybrid mapping of five muscarinic acetylcholine receptor subtype genes

in Rattus norvegicus.

Tseng J(1), Erbe CB, Kwitek AE, Jacob HJ, Popper P, Wackym PA.

Author information:

(1)Department of Otolaryngology and Communication Sciences, Medical College of

Wisconsin, 9200 W. Wisconsin Ave., Milwaukee, WI 53226, USA.

Acetylcholine is the main neurotransmitter of the vestibular efferent system and

a wide variety of muscarinic and nicotinic acetylcholine receptors are expressed

in the vestibular periphery. The role of these receptors and in particular the

role of muscarinic acetylcholine receptors in the physiology of the vestibular

neuroepithelium is not understood. Congenic and consomic rats are a convenient

way to investigate the involvement of candidate genes in the manifestation of

defined traits. To use congenic or consomic rats to elucidate the roles of these

receptors in vestibular physiology or pathology the chromosomal location of the

genes encoding these receptors has to be determined. Using radiation hybrid (RH)

mapping and a rat RH map server (www.rgd.mcw.edu/RHMAP SERVER/), we determined

the chromosomal locations of the muscarinic acetylcholine receptor genes in the

rat (Rattus norvegicus). The m1-m5 muscarinic subtypes mapped to the following

chromosomes: Chrm1, chromosome 1; Chrm2, chromosome 4; Chrm3, chromosome 17;

Chrm4, chromosome 3; and Chrm5, chromosome 3. With the chromosomal location for

each of these muscarinic subtypes known, it is now possible to develop congenic

and consomic strains of rats that can be used to study the functions of each of

these subtypes.

PMID: 12433399 [Indexed for MEDLINE]

3548. J Am Med Inform Assoc. 2002 Nov-Dec;9(6):612-20.

Creating an online dictionary of abbreviations from MEDLINE.

Chang JT(1), Schütze H, Altman RB.

Author information:

(1)Department of Genetics, Stanford Medical Informatics, Stanford, California

94305, USA.

OBJECTIVE: The growth of the biomedical literature presents special challenges

for both human readers and automatic algorithms. One such challenge derives from

the common and uncontrolled use of abbreviations in the literature. Each

additional abbreviation increases the effective size of the vocabulary for a

field. Therefore, to create an automatically generated and maintained lexicon of

abbreviations, we have developed an algorithm to match abbreviations in text with

their expansions.

DESIGN: Our method uses a statistical learning algorithm, logistic regression, to

score abbreviation expansions based on their resemblance to a training set of

human-annotated abbreviations. We applied it to Medstract, a corpus of MEDLINE

abstracts in which abbreviations and their expansions have been manually

annotated. We then ran the algorithm on all abstracts in MEDLINE, creating a

dictionary of biomedical abbreviations. To test the coverage of the database, we

used an independently created list of abbreviations from the China Medical

Tribune.

MEASUREMENTS: We measured the recall and precision of the algorithm in

identifying abbreviations from the Medstract corpus. We also measured the recall

when searching for abbreviations from the China Medical Tribune against the

database.

RESULTS: On the Medstract corpus, our algorithm achieves up to 83% recall at 80%

precision. Applying the algorithm to all of MEDLINE yielded a database of 781,632

high-scoring abbreviations. Of all the abbreviations in the list from the China

Medical Tribune, 88% were in the database.

CONCLUSION: We have developed an algorithm to identify abbreviations from text.

We are making this available as a public abbreviation server at

\url[http://abbreviation.stanford.edu/].

PMCID: PMC349378

PMID: 12386112 [Indexed for MEDLINE]

3549. Nihon Yakurigaku Zasshi. 2002 Nov;120(1):43P-46P.

[Modeling of all genome and database].

[Article in Japanese]

Umeyama H(1).

Author information:

(1)Department of Biomolecular Design, School of Pharmaceutical Sciences,

Kitasato-University, Tokyo 108-8641, Japan.

We have developed the protein modeling software FAMS (Full automatic protein

modeling system), and using the FAMS the proteins coded in the all the genes were

modeled. And we developed web browsing software. We had participated in the

CAFASP2 contest of the CASP4 which is the competition of the protein structure

prediction. We won almost best server in the CAFASP2 which is the contest of full

automatic protein modeling. Accordingly the database quality made by using the

FAMS program will be very good. The FAMS modeling web service is available in

http://physchem.pharm.kitasato-u.ac.jp/. FAMSBASE is seen in the web site of

http://famsbase.bio.nagoya-u.ac.jp/.

PMID: 12491776 [Indexed for MEDLINE]

3550. BMC Bioinformatics. 2002 Oct 25;3:32. Epub 2002 Oct 25.

SeqHound: biological sequence and structure database as a platform for

bioinformatics research.

Michalickova K(1), Bader GD, Dumontier M, Lieu H, Betel D, Isserlin R, Hogue CW.

Author information:

(1)Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada

M5S 1A8. katerina@mshri.on.ca

BACKGROUND: SeqHound has been developed as an integrated biological sequence,

taxonomy, annotation and 3-D structure database system. It provides a

high-performance server platform for bioinformatics research in a locally-hosted

environment.

RESULTS: SeqHound is based on the National Center for Biotechnology Information

data model and programming tools. It offers daily updated contents of all Entrez

sequence databases in addition to 3-D structural data and information about

sequence redundancies, sequence neighbours, taxonomy, complete genomes,

functional annotation including Gene Ontology terms and literature links to

PubMed. SeqHound is accessible via a web server through a Perl, C or C++ remote

API or an optimized local API. It provides functionality necessary to retrieve

specialized subsets of sequences, structures and structural domains. Sequences

may be retrieved in FASTA, GenBank, ASN.1 and XML formats. Structures are

available in ASN.1, XML and PDB formats. Emphasis has been placed on complete

genomes, taxonomy, domain and functional annotation as well as 3-D structural

functionality in the API, while fielded text indexing functionality remains under

development. SeqHound also offers a streamlined WWW interface for simple web-user

queries.

CONCLUSIONS: The system has proven useful in several published bioinformatics

projects such as the BIND database and offers a cost-effective infrastructure for

research. SeqHound will continue to develop and be provided as a service of the

Blueprint Initiative at the Samuel Lunenfeld Research Institute. The source code

and examples are available under the terms of the GNU public license at the

Sourceforge site http://sourceforge.net/projects/slritools/ in the SLRI Toolkit.

PMCID: PMC138791

PMID: 12401134 [Indexed for MEDLINE]

3551. Planta. 2002 Oct;215(6):1031-9. Epub 2002 Jul 25.

Biosynthesis of podophyllotoxin in Linum album cell cultures.

Seidel V(1), Windhövel J, Eaton G, Alfermann AW, Arroo RR, Medarde M, Petersen M,

Woolley JG.

Author information:

(1)Natural Products Research, Faculty of Applied Sciences, De Montfort

University, the Gateway, Leicester LEI 9BH, UK. veronique.seidel@ulsop.ac.uk

Cell cultures of Linum album Kotschy ex Boiss. (Linaceae) showing high

accumulation of the lignan podophyllotoxin (PTOX) were established. Enzymological

studies revealed highest activities of phenylalanine ammonia-lyase, cinnamyl

alcohol dehydrogenase, 4-hydroxycinnamate:CoA ligase and cinnamoyl-CoA:NADP

oxidoreductase immediately prior to PTOX accumulation. To investigate PTOX

biosynthesis, feeding experiments were performed with

[2-(13)C]3',4'-dimethoxycinnamic acid, [2-(13)C]3',4'-methylenedioxycinnamic acid

(MDCA), [2-(13)C]3',4',5'-trimethoxycinnamic acid, [2-(13)C]sinapic acid,

[2-(13)C]- and [2,3-(13)C(2)]ferulic acid. Analysis of the metabolites by HPLC

coupled to tandem mass spectrometry revealed incorporation of label from ferulic

acid into PTOX and deoxypodophyllotoxin (DOP). In addition, MDCA was also

unambiguously incorporated intact into PTOX. These observations suggest that in

L. album both ferulic acid and methylenedioxy-substituted cinnamic acid can be

incorporated into lignans. Furthermore, it appears that, in this species, the

hydroxylation of DOP is a rate-limiting point in the pathway leading to PTOX.

Electronic supplementary material to this paper can be obtained by using the

Springer LINK server located at http://dx.doi.org/wo.1007/s00425-002-0834-1.

DOI: 10.1007/s00425-002-0834-1

PMID: 12355164 [Indexed for MEDLINE]

3552. Planta. 2002 Oct;215(6):914-23. Epub 2002 Jul 25.

Nitric oxide induces transcriptional activation of the nitric oxide-tolerant

alternative oxidase in Arabidopsis suspension cells.

Huang X(1), von Rad U, Durner J.

Author information:

(1)Institute of Biochemical Plant Pathology, GSF - National Research Center for

Environment and Health, 85764 Oberschleissheim, Germany.

Nitric oxide (NO) is a double-edged sword - it can be either beneficial and

activate defence responses in plants and animals or, together with reactive

oxygen species, it can kill not only the pathogen but also the host. A prime

target of NO is the cytochrome c-dependent respiration. Only plants possess

alternative-pathway respiration with alternative oxidase (AOX) as a terminal

electron acceptor. AOX has been suggested to be barely affected by NO. Here we

show that NO affects cytochrome-dependent respiration in Arabidopsis thaliana

(L.) Heynh. At the same time, treatment of Arabidopsis cell cultures with NO

actually strongly induced AOX1a transcription, as determined by using a cDNA

microarray and by Northern analysis. In accordance with transcript accumulation,

NO treatment of suspension cells resulted in increased respiration through the

alternative pathway. Addition of an AOX inhibitor to Arabidopsis cell cultures

resulted in dramatically increased NO-sensitivity and cell death. In all, our

data suggest that NO induces the AOX1a gene and that AOX may participate to

counteract the toxicity of NO. Electronic supplementary material to this paper

can be obtained by using the Springer Link server located at

http://dx.doi.org/10.1007/s00425-002-0828-z.

DOI: 10.1007/s00425-002-0828-z

PMID: 12355151 [Indexed for MEDLINE]

3553. Space Med Med Eng (Beijing). 2002 Oct;15(5):369-73.

[A design of ultrasonic tele-diagnosis and quality control system].

[Article in Chinese]

Wang L, Li DY, Zhao SK, Yin LX, Zhen CQ, Wang TF.

Objective. To study the method for constructing an ultrasonic tele-diagnosis and

quality control system. Method. The standard of Intemet/Intranet technology was

used for the system. The compression and up-loading component based on wavelet

and the software platform of ultrasound image processing were developed to suit

the ultrasound image characteristics. The information exchange and storage

(DICOM) was studied for the compatibility between the system and the Hospital

Information System (HIS). Both of them were combined to form an integrated

system. Result. WWW server, relevant case history managing and image processing

software were constructed in the system. Conclusion. It is feasible to establish

a low cost and maintainable system for ultrasound tele-diagnosis and quality

control system based on Internet/Intranet technology.

PMID: 12449146 [Indexed for MEDLINE]

3554. Bioinformatics. 2002 Sep;18(9):1280-1.

An ontology driven architecture for derived representations of macromolecular

structure.

Greer DS(1), Westbrook JD, Bourne PE.

Author information:

(1)The Research Collaboratory for Structural Bioinformatics at the San Diego

Supercomputer Center, 9500 Gilman Drive, University of California, San Diego, La

Jolla CA 92093-0505, USA. dsg@sdsc.edu

An object metamodel based on a standard scientific ontology has been developed

and used to generate a CORBA interface, an SQL schema and an XML representation

for macromolecular structure (MMS) data. In addition to the interface and schema

definitions, the metamodel was also used to generate the core elements of a CORBA

reference server and a JDBC database loader. The Java source code which

implements this metamodel, the CORBA server, database loader and XML converter

along with detailed documentation and code examples are available as part of the

OpenMMS toolkit.AVAILABILITY: http://openmms.sdsc.edu

CONTACT: dsg@sdsc.edu

PMID: 12217927 [Indexed for MEDLINE]

3555. Hum Immunol. 2002 Sep;63(9):701-9.

Prediction of MHC class I binding peptides using profile motifs.

Reche PA(1), Glutting JP, Reinherz EL.

Author information:

(1)Laboratory of Immunobiology, Dana-Farber Cancer Institute, Boston, MA, USA.

Peptides that bind to a given major histocompatibility complex (MHC) molecule

share sequence similarity. Therefore, a position specific scoring matrix (PSSM)

or profile derived from a set of peptides known to bind to a specific MHC

molecule would be a suitable predictor of whether other peptides might bind, thus

anticipating possible T-cell epitopes within a protein. In this approach, the

binding potential of any peptide sequence (query) to a given MHC molecule is

linked to its similarity to a group of aligned peptides known to bind to that

MHC, and can be obtained by comparing the query to the PSSM. This article

describes the derivation of alignments and profiles from a collection of peptides

known to bind a specific MHC, compatible with the structural and molecular basis

of the peptide-MHC class I (MHCI) interaction. Moreover, in order to apply these

profiles to the prediction of peptide-MHCI binding, we have developed a new

search algorithm (RANKPEP) that ranks all possible peptides from an input protein

using the PSSM coefficients. The predictive power of the method was evaluated by

running RANKPEP on proteins known to bear MHCI K(b)- and D(b)-restricted T-cell

epitopes. Analysis of the results indicates that > 80% of these epitopes are

among the top 2% of scoring peptides. Prediction of peptide-MHC binding using a

variety of MHCI-specific PSSMs is available on line at our RANKPEP web server

(www.mifoundation.org/Tools/rankpep.html). In addition, the RANKPEP server also

allows the user to enter additional profiles, making the server a powerful and

versatile computational biology benchmark for the prediction of peptide-MHC

binding.

PMID: 12175724 [Indexed for MEDLINE]

3556. J Mol Model. 2002 Sep;8(9):266-71. Epub 2002 Sep 4.

Molecular dynamics studies on the aggregation of Y-shaped fluoroalkanes.

Friedemann R(1), Naumann S, Brickmann J.

Author information:

(1)Institut für Organische Chemie, Martin-Luther-Universität Halle-Wittenberg,

Kurt-Mothes-Str. 2, 06120 Halle(Saale), Germany. friedemann@chemie.uni-halle.de

Molecular dynamics (MD) calculations have been performed on the aggregation of

clusters with up to 128 Y-shaped perfluoroalkylated molecules of the type

C10F20[C7H15]2 (Y-A/128) and C10H20[C7F15]2 (Y-B/128) as well as mixed clusters

(Y-A/64+Y-B/64) using the AMBER 5 program. The effect of the segregation tendency

of the chemically different parts and the influence of the steric repulsion due

to the wedge shape of the molecules on the structure formation have been studied.

The results have been analyzed by snapshots, radial atom pair distribution

functions, orientational correlation functions as well as diffusion coefficients

and are compared with the corresponding findings on clusters of alkanes and

perfluoroalkanes. Electronic supplementary material to this paper can be obtained

by using the Springer LINK server located at

http://dx.doi.org/10.1007/s008940020092y.

DOI: 10.1007/s00894-002-0092-y

PMID: 12415331 [Indexed for MEDLINE]

3557. Int J Health Geogr. 2002 Aug 9;1(1):1.

A simple method for serving Web hypermaps with dynamic database drill-down.

Boulos MN(1), Roudsari AV, Carson ER.

Author information:

(1)Centre for Measurement and Information in Medicine, School of Informatics,

City University, London EC1V 0HB, UK. M.Nabih-Kamel-Boulos@city.ac.uk

BACKGROUND: HealthCyberMap http://healthcybermap.semanticweb.org aims at mapping

parts of health information cyberspace in novel ways to deliver a semantically

superior user experience. This is achieved through "intelligent" categorisation

and interactive hypermedia visualisation of health resources using metadata,

clinical codes and GIS. HealthCyberMap is an ArcView 3.1 project. WebView, the

Internet extension to ArcView, publishes HealthCyberMap ArcView Views as Web

client-side imagemaps. The basic WebView set-up does not support any GIS database

connection, and published Web maps become disconnected from the original project.

A dedicated Internet map server would be the best way to serve HealthCyberMap

database-driven interactive Web maps, but is an expensive and complex solution to

acquire, run and maintain. This paper describes HealthCyberMap simple, low-cost

method for "patching" WebView to serve hypermaps with dynamic database drill-down

functionality on the Web. RESULTS: The proposed solution is currently used for

publishing HealthCyberMap GIS-generated navigational information maps on the Web

while maintaining their links with the underlying resource metadata base.

CONCLUSION: The authors believe their map serving approach as adopted in

HealthCyberMap has been very successful, especially in cases when only map

attribute data change without a corresponding effect on map appearance. It should

be also possible to use the same solution to publish other interactive GIS-driven

maps on the Web, e.g., maps of real world health problems.

PMCID: PMC131013

PMID: 12437788

3558. Bioinformatics. 2002 Aug;18(8):1149-50.

The EBI SRS server-new features.

Zdobnov EM(1), Lopez R, Apweiler R, Etzold T.

Author information:

(1)EMBL Outstation - European Bioinformatics Institute, Wellcome Trust Genome

Campus, Hinxton, Cambridge, CB10 1SD, UK.

MOTIVATION: Here we report on recent developments at the EBI SRS server

(http://srs.ebi.ac.uk). SRS has become an integration system for both data

retrieval and sequence analysis applications. The EBI SRS server is a primary

gateway to major databases in the field of molecular biology produced and

supported at EBI as well as European public access point to the MEDLINE database

provided by US National Library of Medicine (NLM). It is a reference server for

latest developments in data and application integration. The new additions

include: concept of virtual databases, integration of XML databases like the

Integrated Resource of Protein Domains and Functional Sites (InterPro), Gene

Ontology (GO), MEDLINE, Metabolic pathways, etc., user friendly data

representation in 'Nice views', SRSQuickSearch bookmarklets.

AVAILABILITY: SRS6 is a licensed product of LION Bioscience AG freely available

for academics. The EBI SRS server (http://srs.ebi.ac.uk) is a free central

resource for molecular biology data as well as a reference server for the latest

developments in data integration.

PMID: 12176845 [Indexed for MEDLINE]

3559. J Mol Model. 2002 Aug;8(8):246-52.

Theoretical investigation of the structure and acid-base properties of potential

2-thiolumazine tautomeric forms using the AM1 semiempirical method.

Acuña-Cueva ER(1), Jiménez-Pulido SB, Moreno-Carretero MN.

Author information:

(1)Departamento de Química Inorgánica y Orgánica, Universidad de Jaén,

23071-Jaén, Spain.

The relative stabilities of potential tautomers, in both gas and aqueous phases,

have been calculated taking into account the entropy effects over the tautomeric

equilibria, in order to determine the structure and acid-base properties of the

most stable tautomers of 2-thiolumazine in different pH conditions. In each

medium, the tautomer with the lowest energy must be the most representative form

at the corresponding pH. Knowledge of the effect of the medium in the

tautomerization energies allows us to evaluate the possible effect of the medium

on the molecular stability. Clearly, the results show that in the gas phase the

basicity of the potential donor atoms is N5<N8<O4<S2<N1<N3, and in the aqueous

phase S2<(O4 approximately N5)<N8<N1<N3, with the higher basicity of N3 and N1

being common to the two phases. In the aqueous phase, the sulfur atom is usually

found in the thiol form, whereas the oxygen atom is in the keto form only in the

most stable species. Moreover the acid-base character of 2-thiolumazine in

aqueous solution has been evaluated from the corresponding AM1 thermodynamic

parameters. The results agree well with the experimental data. Electronic

supplementary material to this paper can be obtained by using the Springer Link

server located at http://dx.doi.org/10.1007/s00894-002-0094-9

DOI: 10.1007/s00894-002-0094-9

PMID: 12324801 [Indexed for MEDLINE]

3560. Nat Biotechnol. 2002 Aug;20(8):835-9. Epub 2002 Jul 8.

An algorithm for finding protein-DNA binding sites with applications to

chromatin-immunoprecipitation microarray experiments.

Liu XS(1), Brutlag DL, Liu JS.

Author information:

(1)Stanford Medical Informatics, Stanford University, Stanford CA 94305, USA.

Chromatin immunoprecipitation followed by cDNA microarray hybridization

(ChIP-array) has become a popular procedure for studying genome-wide protein-DNA

interactions and transcription regulation. However, it can only map the probable

protein-DNA interaction loci within 1-2 kilobases resolution. To pinpoint

interaction sites down to the base-pair level, we introduce a computational

method, Motif Discovery scan (MDscan), that examines the ChIP-array-selected

sequences and searches for DNA sequence motifs representing the protein-DNA

interaction sites. MDscan combines the advantages of two widely adopted motif

search strategies, word enumeration and position-specific weight matrix updating,

and incorporates the ChIP-array ranking information to accelerate searches and

enhance their success rates. MDscan correctly identified all the experimentally

verified motifs from published ChIP-array experiments in yeast (STE12, GAL4,

RAP1, SCB, MCB, MCM1, SFF, and SWI5), and predicted two motif patterns for the

differential binding of Rap1 protein in telomere regions. In our studies, the

method was faster and more accurate than several established motif-finding

algorithms. MDscan can be used to find DNA motifs not only in ChIP-array

experiments but also in other experiments in which a subgroup of the sequences

can be inferred to contain relatively abundant motif sites. The MDscan web server

can be accessed at http://BioProspector.stanford.edu/MDscan/.

DOI: 10.1038/nbt717

PMID: 12101404 [Indexed for MEDLINE]

3561. Naturwissenschaften. 2002 Aug;89(8):352-6. Epub 2002 Jun 22.

Vocal acrobatics in a Chinese frog, Amolops tormotus.

Feng AS(1), Narins PM, Xu CH.

Author information:

(1)Department of Molecular and Integrative Physiology, University of Illinois,

Urbana 61801, USA. a-feng@uiuc.edu

Although amphibians are highly vocal, they generally emit only a limited number

of acoustic communication signals. We report here the extraordinarily rich vocal

repertoire of Amolops tormotus, a ranid species in China. These frogs produce

countless vocalizations, some of which share features of birdsong or primate

calls, e.g., ultrasonic frequency components, multiple upward and downward FM

sweeps, and sudden onset and offset of selective harmonic components within a

call note. Frame-by-frame video analysis of the frog's calling behavior suggests

the presence of two pairs of vocal sacs that may contribute to the remarkable

call-note complexity in this species. Electronic supplementary material to this

paper can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00114-002-0335-x.

DOI: 10.1007/s00114-002-0335-x

PMID: 12435035 [Indexed for MEDLINE]

3562. Eur J Med Res. 2002 Jul 24;7(7):323-9.

Radata - implementation of resistance analysis and expert advice for optimized

HAART switches in general practice of HIV-infected individuals via a compiling

internet presence.

Paech V(1), Lorenzen T, Stoehr A, Plettenberg A.

Author information:

(1)Ifi-Institute for interdisciplinary Infectiology and Immunology GmbH, General

Hospital St. Georg, Hamburg, Germany.

BACKGROUND: HIV-infected patients fail viral load suppression, because resistance

against antiretroviral drugs arises or for other reasons. HIV-resistance analyses

can aid to achieve effective HAART regimen. Furthermore, clinical benefits from

genotyping in study settings are significantly higher for treating physicians,

who can include external advice from HIV-experts into HAART switch.

OBJECTIVE: To develop a compiling internet presence to provide expert advice for

HAART switch in general practice of HIV-infected individuals after therapy

failure. -

DESIGN: A multifactorial (genotyping, drug monitoring, adherence, expert advice)

interdisciplinary internet service (www.radata.de) with an associated server

hosted database.

PATIENTS AND METHODS: HIV-infected patients after failure to HAART are eligible

for registration to the Radata project. Genotyping is performed according to

protocols specific for each participating institution. Therapeutic drug

monitoring (NNRTIs, PIs) follows setting for drug level detection by mass

spectrometry. An adherence self-report is completed by every patient. Clinical

documentation is provided by the treating Primary Care Physician. Clinical expert

advice for implementation into HAART switch in daily clinical practice for

treating physicians is provided by HIV-experts according to data obtained.

Clinical and laboratory follow-up visits are scheduled firstly 4 weeks after

HAART switch and three monthly afterwards, over a period of one year.

RESULTS: Technical resources and a compiling internet presence for generation of

resistance analysis based expert advice were developed. Initially, 7

HIV-treatment centres, 7 laboratories and 17 HIV advisors contribute to Radata

database project. 15 patients were enrolled during test period. 30 expert advices

were generated during the test phase. Expert advice was provided in 6 weeks

median for implementation into HAART switch. 13 out of 15 expert advices were

implemented into HAART switch by treating Primary Care Physicians.

CONCLUSIONS: Radata is a novel database concept with features to generate expert

advice for implementation into HAART switch of HIV-infected subjects. A test

period has shown, that the concept is technically approved to fit all

requirements with regard to data collection, evaluation and to generate expert

advice for therapy switch in daily clinical practice.

PMID: 12176682 [Indexed for MEDLINE]

3563. Anal Bioanal Chem. 2002 Jul;373(4-5):266-76. Epub 2002 Jun 6.

Quick measurement of protein sulfhydryls with Ellman's reagent and with

4,4'-dithiodipyridine.

Riener CK(1), Kada G, Gruber HJ.

Author information:

(1)Institute of Biophysics, J. Kepler University, Altenberger Str. 69, A-4040

Linz, Austria.

Since its introduction in 1959, Ellman's reagent (5,5'-dithio-bis(2-nitrobenzoic

acid)) has been the favorite reagent for spectrophotometric measurement of

protein sulfhydryls. Meanwhile however, evidence has accumulated that many

protein sulfhydryls give an incomplete reaction with Ellman's reagent, even

during prolonged assay times. In the present study, the kinetic problem was

solved by including cystamine as a "mediator" between the protein sulfhydryl and

Ellman's reagent, as previously applied in an enzymatic thiol assay [9]. As an

alternative, 4,4'-dithiodipyridine (DTDP) was used in place of Ellman's reagent.

Due to its small size, amphiphilic nature, and lack of charge, DTDP quickly

reacts with poorly accessible protein sulfhydryls, without any catalysis by

cystamine. The DTDP method and the Ellman/cystamine method were both optimized

for maximal sensitivity, minimal sample consumption (detection limit 0.2 nmol

mL(-1), determination limit 0.6 nmol mL(-1)), and minimal assay time (5 min). In

validation experiments, both methods gave identical results and the measured

sulfhydryls/protein matched the expected values. Electronic supplementary

material to this paper can be obtained by using the Springer Link server located

at http://dx.doi.org/10.1007/s00216-002-1347-2.

DOI: 10.1007/s00216-002-1347-2

PMID: 12110978 [Indexed for MEDLINE]

3564. Arch Microbiol. 2002 Jul;178(1):65-70. Epub 2002 Apr 30.

Aciduric Proteobacteria isolated from pH 2.9 soil.

Curtis P(1), Nakatsu CH, Konopka A.

Author information:

(1)Department of Biological Sciences, Purdue University, West Lafayette IN 47907,

USA.

Acidic (pH 2.9) soil was used as an inoculum to culture heterotrophic bacteria at

pH values of 3-4. Four isolates were obtained; on the basis of 16S rDNA sequence,

they were shown to be members of the beta- and gamma-Proteobacteria. The three

isolates that were most closely related to Burkholderia spp. had simple

nutritional requirements and could grow in glucose-mineral salts media; two of

these used a broad array of organic substrates. The 16S rDNA sequence of the

fourth isolate was most similar (96%) to Frateuria aurantia. The isolates were

aciduric rather than acidophilic; their pH ranges for growth were approximately

3.5-8. Unlike many bacteria whose acid tolerance represents the capacity to

survive acid exposure, these microorganisms carried out exponential growth at

pH<4 and their growth rates at pH 3.9 ranged from 60 to 98% of those found at pH

7. The cell yields on glucose of two strains were identical at pH 4 and pH 7. The

acidic soils appeared to contain a very diverse bacterial community as assessed

by denaturing gradient gel electrophoresis fingerprinting of PCR amplicons of a

portion of the 16S rDNA gene. Electronic supplementary material to this paper can

be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00203-002-0427-1.

DOI: 10.1007/s00203-002-0427-1

PMID: 12070771 [Indexed for MEDLINE]

3565. Bioinformatics. 2002 Jul;18(7):1017-9.

SNAPper: gene order predicts gene function.

Kolesov G(1), Mewes HW, Frishman D.

Author information:

(1)Institute for Bioinformatics, GSF-National Research Center for Environment and

Heath, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany. g.kolesov@gsf.de

SNAPper is a network service for predicting gene function based on the

conservation of gene order.AVAILABILITY: The SNAPper server is available at

http://pedant.gsf.de/snapper. SNAPper-based functional predictions will soon be

offered as part of the PEDANT genome analysis server http://pedant.gsf.de.

PMID: 12117803 [Indexed for MEDLINE]

3566. Eur Biophys J. 2002 Jul;31(4):317-22. Epub 2002 Jun 14.

Lateral organization of GM1 in phase-separated monolayers visualized by scanning

force microscopy.

Menke M(1), Künneke S, Janshoff A.

Author information:

(1)Institut für Physikalische Chemie, Welder Weg 11, Johannes

Gutenberg-Universität, 55099 Mainz, Germany.

Phase separation of glycolipids in lipid mono- and bilayers is of great interest

for the understanding of membrane function. The distribution of the ganglioside

GM1 in sphingomyelin (SM)/1-palmitoyl-2-oleoyl- sn-glycero-3-phosphocholine

(POPC), SM/1,2-dipalmitoyl- sn-glycero-3-phosphocholine (DOPC) and

SM/cholesterol/POPC Langmuir-Blodgett (LB) monolayers transferred at 36 mN/m has

been studied by scanning force microscopy. Besides lateral organization of the

glycolipid in LB monolayers as deduced from topography, material properties have

been investigated by phase imaging, pulsed force mode and force modulation

microscopy. It was shown that GM1 preferentially clusters in an ordered lipid

matrix, i.e. the SM phase in the case of the SM/POPC and SM/DOPC mixture or in

the ordered phase of POPC/SM/cholesterol monolayers. At higher local

concentrations, three-dimensional protrusions enriched in GM1 occur, which may

represent a precursor for the formation of micelles budding into the aqueous

subphase. Electronic supplementary material to this paper can be obtained by

using the Springer Link server located at

http://dx.doi.org/10.1007/s00249-002-0232-4.

DOI: 10.1007/s00249-002-0232-4

PMID: 12122478 [Indexed for MEDLINE]

3567. Med Sci Monit. 2002 Jul;8(7):MT124-36.

Towards a semantic medical Web: HealthCyberMap's tool for building an RDF

metadata base of health information resources based on the Qualified Dublin Core

Metadata Set.

Boulos MN(1), Roudsari AV, Carson ER.

Author information:

(1)Centre for Measurement and Information in Medicine, School of Informatics,

City University, London, UK. M.Nabih-Kamel-Boulos@city.ac.uk

BACKGROUND: HealthCyberMap (http://healthcybermap.semanticweb.org/) aims at

mapping Internet health information resources in novel ways for enhanced

retrieval and navigation. This is achieved by collecting appropriate resource

metadata in an unambiguous form that preserves semantics.

MATERIAL/METHODS: We modelled a qualified Dublin Core (DC) metadata set ontology

with extra elements for resource quality and geographical provenance in Prot g

-2000. A metadata collection form helps acquiring resource instance data within

Prot g . The DC subject field is populated with UMLS terms directly imported from

UMLS Knowledge Source Server using UMLS tab, a Prot g -2000 plug-in. The project

is saved in RDFS/RDF.

RESULTS: The ontology and associated form serve as a free tool for building and

maintaining an RDF medical resource metadata base. The UMLS tab enables browsing

and searching for concepts that best describe a resource, and importing them to

DC subject fields. The resultant metadata base can be used with a search and

inference engine, and have textual and/or visual navigation interface(s) applied

to it, to ultimately build a medical Semantic Web portal. Different ways of

exploiting Prot g -2000 RDF output are discussed.

CONCLUSIONS: By making the context and semantics of resources, not merely their

raw text and formatting, amenable to computer 'understanding,' we can build a

Semantic Web that is more useful to humans than the current Web. This requires

proper use of metadata and ontologies. Clinical codes can reliably describe the

subjects of medical resources, establish the semantic relationships (as defined

by underlying coding scheme) between related resources, and automate their

topical categorisation.

PMID: 12118210 [Indexed for MEDLINE]

3568. Bioinformatics. 2002 Jun;18(6):819-24.

NETASA: neural network based prediction of solvent accessibility.

Ahmad S(1), Gromiha MM.

Author information:

(1)Institute of Multimedia and Software, Universiti Putra Malaysia, Serdang,

43400, Selangor, Malaysia. shandar@jamia.net

MOTIVATION: Prediction of the tertiary structure of a protein from its amino acid

sequence is one of the most important problems in molecular biology. The

successful prediction of solvent accessibility will be very helpful to achieve

this goal. In the present work, we have implemented a server, NETASA for

predicting solvent accessibility of amino acids using our newly optimized neural

network algorithm. Several new features in the neural network architecture and

training method have been introduced, and the network learns faster to provide

accuracy values, which are comparable or better than other methods of ASA

prediction.

RESULTS: Prediction in two and three state classification systems with several

thresholds are provided. Our prediction method achieved the accuracy level upto

90% for training and 88% for test data sets. Three state prediction results

provide a maximum 65% accuracy for training and 63% for the test data.

Applicability of neural networks for ASA prediction has been confirmed with a

larger data set and wider range of state thresholds. Salient differences between

a linear and exponential network for ASA prediction have been analysed.

AVAILABILITY: Online predictions are freely available at: http://www.netasa.org.

Linux ix86 binaries of the program written for this work may be obtained by email

from the corresponding author.

PMID: 12075017 [Indexed for MEDLINE]

3569. Eur Biophys J. 2002 Jun;31(3):172-8. Epub 2002 Jan 29.

Sampling the conformational space of membrane protein surfaces with the AFM.

Scheuring S(1), Müller DJ, Stahlberg H, Engel HA, Engel A.

Author information:

(1)M.E. Müller Institute for Microscopy at the Biozentrum, University of Basel,

Klingelbergstrasse 70, 4056 Basel, Switzerland.

The atomic force microscope acquires topographs of single native membrane

proteins at subnanometer resolution. Owing to the high signal-to-noise ratio,

such images allow the conformational space of membrane protein surfaces to be

sampled. This is demonstrated by topographs of porin OmpF, aquaporin-Z, and

bacteriorhodopsin, all recorded at a lateral resolution of <7 A and a vertical

resolution of ~1 A. The amplitudes of the domain movements were estimated from a

large number of single molecule topographs and the corresponding energy

landscapes calculated. To visualize the motion of protein domains, movies were

generated by similarity ranking of the observed protein configurations.

Electronic supplementary material to this paper can be obtained by using the

Springer Link server located at http://dx.doi.org/10.1007/s00249-001-0197-8

DOI: 10.1007/s00249-001-0197-8

PMID: 12029329 [Indexed for MEDLINE]

3570. Hum Genet. 2002 Jun;110(6):592-600. Epub 2002 May 22.

Homogeneity and distinctiveness of Polish paternal lineages revealed by Y

chromosome microsatellite haplotype analysis.

Ploski R(1), Wozniak M, Pawlowski R, Monies DM, Branicki W, Kupiec T, Kloosterman

A, Dobosz T, Bosch E, Nowak M, Lessig R, Jobling MA, Roewer L, Kayser M.

Author information:

(1)Human Molecular Genetics Lab, Department of Forensic Medicine, Warsaw Medical

Academy, Poland.

Different regional populations from Poland were studied in order to assess the

genetic heterogeneity within Poland, investigate the genetic relationships with

other European populations and provide a population-specific reference database

for anthropological and forensic studies. Nine Y-chromosomal microsatellites were

analysed in a total of 919 unrelated males from six regions of Poland and in

1,273 male individuals from nine other European populations. AMOVA revealed that

all of the molecular variation in the Polish dataset is due to variation within

populations, and no variation was detected among populations of different regions

of Poland. However, in the non-Polish European dataset 9.3% ( P<0.0001) of the

total variation was due to differences among populations. Consequently,

differences in R(ST)-values between all possible pairs of Polish populations were

not statistically significant, whereas significant differences were observed in

nearly all comparisons of Polish and non-Polish European populations.

Phylogenetic analyses demonstrated tight clustering of Polish populations

separated from non-Polish groups. Population clustering based on Y-STR haplotypes

generally correlates well with the geography and history of the region. Thus, our

data are consistent with the assumption of homogeneity of present-day paternal

lineages within Poland and their distinctiveness from other parts of Europe, at

least in respect to their Y-STR haplotypes. Electronic supplementary material to

this paper can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00439-002-0728-0.

DOI: 10.1007/s00439-002-0728-0

PMID: 12107446 [Indexed for MEDLINE]

3571. J Digit Imaging. 2002 Jun;15(2):91-7. Epub 2002 Sep 26.

OpenRIMS: an open architecture radiology informatics management system.

Langer S(1).

Author information:

(1)Department of Radiology, University of Washington Medical Center, Seattle, WA,

USA. sglanger@ppsa.com

The benefits of an integrated picture archiving and communication

system/radiology information system (PACS/RIS) archive built with open source

tools and methods are 2-fold. Open source permits an inexpensive development

model where interfaces can be updated as needed, and the code is peer reviewed by

many eyes (analogous to the scientific model). Integration of PACS/RIS

functionality reduces the risk of inconsistent data by reducing interfaces among

databases that contain largely redundant information. Also, wide adoption would

promote standard data mining tools--reducing user needs to learn multiple methods

to perform the same task. A model has been constructed capable of accepting HL7

orders, performing examination and resource scheduling, providing digital imaging

and communications in medicine (DICOM) worklist information to modalities,

archiving studies, and supporting DICOM query/retrieve from third party viewing

software. The multitiered architecture uses a single database communicating via

an ODBC bridge to a Linux server with HL7, DICOM, and HTTP connections. Human

interaction is supported via a web browser, whereas automated informatics

services communicate over the HL7 and DICOM links. The system is still under

development, but the primary database schema is complete as well as key pieces of

the web user interface. Additional work is needed on the DICOM/HL7 interface

broker and completion of the base DICOM service classes.

DOI: 10.1007/s10278-002-0010-y

PMCID: PMC3611608

PMID: 12297975 [Indexed for MEDLINE]

3572. BMC Bioinformatics. 2002 May 16;3:14.

RIO: analyzing proteomes by automated phylogenomics using resampled inference of

orthologs.

Zmasek CM(1), Eddy SR.

Author information:

(1)Howard Hughes Medical Institute and Department of Genetics, Washington

University School of Medicine, St Louis, MO 63110, USA. zmasek@genetics.wustl.edu

BACKGROUND: When analyzing protein sequences using sequence similarity searches,

orthologous sequences (that diverged by speciation) are more reliable predictors

of a new protein's function than paralogous sequences (that diverged by gene

duplication). The utility of phylogenetic information in high-throughput genome

annotation ("phylogenomics") is widely recognized, but existing approaches are

either manual or not explicitly based on phylogenetic trees.

RESULTS: Here we present RIO (Resampled Inference of Orthologs), a procedure for

automated phylogenomics using explicit phylogenetic inference. RIO analyses are

performed over bootstrap resampled phylogenetic trees to estimate the reliability

of orthology assignments. We also introduce supplementary concepts that are

helpful for functional inference. RIO has been implemented as Perl pipeline

connecting several C and Java programs. It is available at

http://www.genetics.wustl.edu/eddy/forester/. A web server is at

http://www.rio.wustl.edu/. RIO was tested on the Arabidopsis thaliana and

Caenorhabditis elegans proteomes.

CONCLUSION: The RIO procedure is particularly useful for the automated detection

of first representatives of novel protein subfamilies. We also describe how some

orthologies can be misleading for functional inference.

PMCID: PMC116988

PMID: 12028595 [Indexed for MEDLINE]

3573. Proteins. 2002 May 15;47(3):393-402.

Increasing the precision of comparative models with YASARA NOVA--a

self-parameterizing force field.

Krieger E(1), Koraimann G, Vriend G.

Author information:

(1)Center for Molecular and Biomolecular Informatics, Nijmegen, The Netherlands.

elmar.krieger@cmbi.kun.nl

One of the conclusions drawn at the CASP4 meeting in Asilomar was that applying

various force fields during refinement of template-based models tends to move

predictions in the wrong direction, away from the experimentally determined

coordinates. We have derived an all-atom force field aimed at protein and

nucleotide optimization in vacuo (NOVA), which has been specifically designed to

avoid this problem. NOVA resembles common molecular dynamics force fields but has

been automatically parameterized with two major goals: (i) not to make high

resolution X-ray structures worse and (ii) to improve homology models built by

WHAT IF. Force-field parameters were not required to be physically correct;

instead, they were optimized with random Monte Carlo moves in force-field

parameter space, each one evaluated by simulated annealing runs of a 50-protein

optimization set. Errors inherent to the approximate force-field equation could

thus be canceled by errors in force-field parameters. Compared with the

optimization set, the force field did equally well on an independent validation

set and is shown to move in silico models closer to reality. It can be applied to

modeling applications as well as X-ray and NMR structure refinement. A new method

to assign force-field parameters based on molecular trees is also presented. A

NOVA server is freely accessible at http://www.yasara.com/servers

Copyright 2002 Wiley-Liss, Inc.

PMID: 11948792 [Indexed for MEDLINE]

3574. Behav Res Methods Instrum Comput. 2002 May;34(2):234-40.

WEXTOR: a Web-based tool for generating and visualizing experimental designs and

procedures.

Reips UD(1), Neuhaus C.

Author information:

(1)Universität Zürich, Attenhoferstr. 9, CH-8032 Zürich, Switzerland.

ureips@genpsy.unizh.ch

WEXTOR is a Javascript-based experiment generator and teaching tool on the

World-Wide Web that can be used to design laboratory and Web experiments in

guided step-by-step process. It dynamically creates the customized Web pages and

Javascripts needed for the experimental procedure and provides experimenters with

a print-ready visual display of their experimental design. WEXTOR flexibly

supports complete and incomplete factorial designs with between-subjects,

within-subjects, and quasi-experimental factors, as well as mixed designs. The

software implements client-side response time measurement and contains a content

wizard for creating interactive materials, as well as dependent measures

(graphical scales, multiple-choice items, etc.), on the experiment pages.

However, it does not aim to replace a full-fledged HTML editor. Several

methodological features specifically needed in Web experimental design have been

implemented in the Web-based tool and are described in this paper. WEXTOR is

platform independent. The created Web pages can be uploaded to any type of Web

server in which data may be recorded in logfiles or via a database. The current

version of WEXTOR is freely available for educational and noncommercial purposes.

Its Web address is http://www.genpsylab.unizh.ch/wextor/index.html.

PMID: 12109018 [Indexed for MEDLINE]

3575. Bioinformatics. 2002 May;18(5):767-8.

ENDscript: a workflow to display sequence and structure information.

Gouet P(1), Courcelle E.

Author information:

(1)Laboratoire de Bio-Cristallographie, IBCP-CNRS UMR 5086, 7 passage du Vercors,

69367 Lyon Cedex 07, France. p.gouet@ibcp.fr

ENDscript is a web server grouping popular programs such as BLAST, Multalin and

DSSP. It uses as query the co-ordinates file of a protein in Protein Data Bank

format and generates PostScript and png figures showing: residues conserved after

a multiple alignment against homologous sequences, secondary structure elements,

accessibility, hydropathy and intermolecular contacts. Thus, the user can relate

quickly 1D, 2D and 3D information of a protein of known structure.AVAILABILITY:

http://genopole.toulouse.inra.fr/ENDscript

PMID: 12050076 [Indexed for MEDLINE]

3576. Bioinformatics. 2002 May;18(5):763-4.

EMBL-Align: a new public nucleotide and amino acid multiple sequence alignment

database.

Lombard V(1), Camon EB, Parkinson HE, Hingamp P, Stoesser G, Redaschi N.

Author information:

(1)EMBL Outstation The European Bioinformatics Institute (EBI), Wellcome Trust

Genome Campus, Hinxton, Cambridge CB10 1SD, UK. align@ebi.ac.uk

The submission of multiple sequence alignment data to EMBL has grown 30-fold in

the past 10 years, creating a problem of archiving them. The EBI has developed a

new public database of multiple sequence alignments called EMBL-Align. It has a

dedicated web-based submission tool, Webin-Align. Together they represent a

comprehensive data management solution for alignment data. Webin-Align accepts

all the common alignment formats and can display data in CLUSTALW format as well

as a new standard EMBL-Align flat file format. The alignments are stored in the

EMBL-Align database and can be queried from the EBI SRS (Sequence Retrieval

System) server.AVAILABILITY: Webin-Align:

http://www.ebi.ac.uk/embl/Submission/align\_top.html, EMBL-Align:

ftp://ftp.ebi.ac.uk/pub/databases/embl/align, http://srs.ebi.ac.uk/

PMID: 12050074 [Indexed for MEDLINE]

3577. Dev Genes Evol. 2002 May;212(4):186-95. Epub 2002 Apr 18.

Characterisation of two snail genes in the gastropod mollusc Patella vulgata.

Implications for understanding the ancestral function of the snail-related genes

in Bilateria.

Lespinet O(1), Nederbragt AJ, Cassan M, Dictus WJ, van Loon AE, Adoutte A.

Author information:

(1)Centre de Génétique Moléculaire, CNRS UPR 2167, 91198 Gif-sur-Yvette, France.

lespinet@ncbi.nlm.nih.gov

Snail genes have been found to play a role in mesoderm formation in two of the

three clades of bilaterians, deuterostomes (comprising the chordates) and

ecdysozoans (comprising the arthropods). No clear data are available on the role

these genes play in development of the mesoderm in the third clade, that of

lophotrochozoans (comprising annelids and molluscs). We identified two new

members of the snail gene family in the gastropod mollusc Patella vulgata.

Phylogenetic analysis showed that the two genes clearly belong to the snail

sub-family. Their expression patterns do not indicate a role during early

mesoderm formation. In fact, contrary to expectations, the snail genes of Patella

were mostly expressed in the ectoderm. In view of the location of their

expression sites, we suggest that these genes could be involved in regulating

epithelial-mesenchymal transitions (EMT) and cell motility, as has recently been

demonstrated for snail genes in vertebrates. This may well correspond to the

ancestral function of these genes. The results are discussed in the light of the

evolutionary origin of the mesoderm. Electronic supplementary material to this

paper can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00427-002-0228-1.

DOI: 10.1007/s00427-002-0228-1

PMID: 12012233 [Indexed for MEDLINE]

3578. Funct Integr Genomics. 2002 May;2(1-2):40-50. Epub 2002 Apr 9.

Two haplotypes of resistance gene analogs have been conserved during evolution at

the leaf rust resistance locus Lr10in wild and cultivated wheat.

Scherrer B(1), Keller B, Feuillet C.

Author information:

(1)Institute of Plant Biology, University of Zürich, Zollikerstrasse 107, 8008

Zürich, Switzerland.

The isolation of genes of agronomic interest such as disease resistance genes is

a central issue in wheat research. A good knowledge of the organization and

evolution of the genome can greatly help in defining the best strategies for

efficient gene isolation. So far, very few wheat disease resistance loci have

been studied at the molecular level and little is known about their evolution

during polyploidization and domestication. In this study, we have analyzed the

haplotype structure at loci orthologous to the leaf rust resistance locus Lr10in

hexaploid wheat which spans 350 kb in diploid wheat. Two haplotypes (H1, H2) were

defined by the presence (H1) or the absence (H2) of two different resistance gene

analogs ( rga1, rga2) at this locus on chromosome 1AS. Both haplotypes were found

in a collection of 113 wild and cultivated diploid and polyploid wheat lines and

they do not reflect phylogenetic relationships. This indicates an ancient origin

for this disease resistance locus and the independent conservation of the two

haplotypes throughout the evolution of the wheat genome. Finally, the coding

regions of the H1 haplotype RGAs are extremely conserved in all the species. This

suggests a selective pressure for maintaining the structural and functional

configuration of this haplotype in wheat. Electronic supplementary material to

this paper can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s10142-002-0051-9.

DOI: 10.1007/s10142-002-0051-9

PMID: 12021849 [Indexed for MEDLINE]

3579. Funct Integr Genomics. 2002 May;2(1-2):13-27. Epub 2002 Feb 19.

DNA array profiling of gene expression changes during maize embryo development.

Lee JM(1), Williams ME, Tingey SV, Rafalski JA.

Author information:

(1)DuPont Crop Genetics, Delaware Technology Park, Newark, 19714, USA..

We are using DNA microarray-based gene expression profiling to classify temporal

patterns of gene expression during the development of maize embryos, to

understand mRNA-level control of embryogenesis and to dissect metabolic pathways

and their interactions in the maize embryo. Genes involved in carbohydrate, fatty

acid, and amino acid metabolism, the tricarboxylic acid (TCA) cycle, glycolysis,

the pentose phosphate pathway, embryogenesis, membrane transport, signal

transduction, cofactor biosynthesis, photosynthesis, oxidative phosphorylation

and electron transfer, as well as 600 random complementary DNA (cDNA) clones from

maize embryos, were arrayed on glass slides. DNA arrays were hybridized with

fluorescent dye-labeled cDNA probes synthesized from kernel and embryo

poly(A)(+)RNA from different stages of maize seed development. Several

characteristic developmental patterns of expression were identified and

correlated with gene function. Patterns of coordinated gene expression in the TCA

cycle and glycolysis were analyzed in detail. The steady state level of

poly(A)(+) RNA for many genes varies dramatically during maize embryo

development. Expression patterns of genes coding for enzymes of fatty acid

biosynthesis and glycolysis are coordinately regulated during development. Genes

of unknown function may by assigned a hypothetical role based on their patterns

of expression resembling well characterized genes. Electronic supplementary

material to this paper can be obtained by using the Springer LINK server located

at http://dx.doi.org/10.1007/s10142-002-0046-6.

DOI: 10.1007/s10142-002-0046-6

PMID: 12021847 [Indexed for MEDLINE]

3580. Intensive Care Med. 2002 May;28(5):656-9. Epub 2002 Apr 13.

Quality of data collected for severity of illness scores in the Dutch National

Intensive Care Evaluation (NICE) registry.

Arts D(1), de Keizer N, Scheffer GJ, de Jonge E.

Author information:

(1)Department of Medical Informatics and Department of Intensive Care, Academic

Medical Center, P.O.Box 22700, 1100 DE Amsterdam, The Netherlands.

D.G.Arts@amc.uva.nl

OBJECTIVE: To analyse the quality of data used to measure severity of illness in

the Dutch National Intensive Care Evaluation (NICE) registry, after

implementation of quality improving procedures.

DESIGN: Data were re-abstracted from the paper records of patients or the Patient

Data Management System and compared to the data contained in the registry. The

re-abstracted data were considered to be the gold standard.

SETTING: ICUs of nine Dutch hospitals that had been collecting data for the NICE

registry for at least 1 year.

MEASUREMENT AND RESULTS: The mean percentages of inaccurate and incomplete data,

per hospital, over all variables, were 6.1%+/-4.4 (SD) and 2.7%+/-4.4 (SD),

respectively. The mean difference in severity of illness scores between registry

data and re-abstracted data was 0.2 points for APACHE II and 0.4 points for SAPS

II. The mean difference in predicted mortality according to APACHE II and SAPS II

between registry data and re-abstracted data was 0.4% and 0.02%, respectively.

CONCLUSIONS: The current data quality of the NICE registry is good and justifies

evaluative research. These positive results might be explained by the

implementation of several quality assurance procedures in the NICE registry, such

as training and automatic data checks. Electronic supplementary material to this

paper can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00134-002-1272-z

DOI: 10.1007/s00134-002-1272-z

PMID: 12029418 [Indexed for MEDLINE]

3581. Mol Genet Genomics. 2002 May;267(3):359-69. Epub 2002 Apr 26.

Combinatorial variation in coding and promoter sequences of genes at the Tri

locus in Pisum sativum accounts for variation in trypsin inhibitor activity in

seeds.

Page D(1), Aubert G, Duc G, Welham T, Domoney C.

Author information:

(1)Laboratoire des Légumineuses, INRA-URGAP Dijon, Domaine d'Epoisses, BP 86510,

21065 Dijon, France.

Cultivars of Pisum sativum that differ with respect to the quantitative

expression of trypsin/chymotrypsin inhibitor proteins in seeds have been examined

in terms of the structure of the corresponding genes. The patterns of divergence

in the promoter and coding sequences are described, and the divergence among

these exploited for the development of facile DNA-based assays to distinguish

genotypes. Quantitative effects on gene expression may be attributed to the

overall gene complement and to particular promoter/coding sequence combinations,

as well as to the existence of distinct active-site variants that ultimately

influence protein activity. Electronic supplementary material to this paper can

be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00438-002-0667-4.

DOI: 10.1007/s00438-002-0667-4

PMID: 12073038 [Indexed for MEDLINE]

3582. Naturwissenschaften. 2002 May;89(5):201-13.

Review of the reticulated python (Python reticulatus Schneider, 1801) with the

description of new subspecies from Indonesia.

Auliya M(1), Mausfeld P, Schmitz A, Böhme W.

Author information:

(1)Zoologisches Forschungsinstitut und Museum Alexander Koenig, Section of

Herpetology, Bonn, Germany. m.auliya.zfmk@uni-bonn.de

The geographically widespread Python reticulatus, the world's longest snake, has

been largely neglected by taxonomists. Dwarfed individuals from Tanahjampea

Island, Indonesia, differ strikingly in morphology. Phylogenetic relationships

were analyzed using a 345-bp fragment of the cytochrome b gene for 12 specimens

from different populations. Both genetic differences and morphological characters

distinctly revealed two taxonomic subunits. The island populations of Tanahjampea

and Selayar form two monophyletic lineages, supported by high bootstrap values,

with distinct differences in color pattern and scalation. We consider these forms

to represent two new subspecies. The Tanahjampea form is genetically related to

populations of the Sunda Islands and mainland Southeast Asia, whereas the Selayar

form is related to populations of Southwest Sulawesi. We conclude that, due to

strong directional surface currents in this region, gene flow between Tanahjampea

and Selayar is prevented. Sea-level changes during the Pleistocene probably

contributed to the isolation of the two taxa described. Aspects of ecology and

conservation status are briefly discussed. Electronic supplementary material to

this paper can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00114-002-0320-4.

PMID: 12135085 [Indexed for MEDLINE]

3583. Surg Radiol Anat. 2002 May;24(2):87-90.

The cutaneo-lymph node flap of the superficial circumflex artery.

Assouad J(1), Becker C, Hidden G, Riquet M.

Author information:

(1)Service de Chirurgie Thoracique, Hôpital Européen Georges Pompidou, 20-40, rue

Leblanc, 75015 Paris, France.

Upper limb lymphoedema and associated radiation-damaged chest wall are

complications occurring after breast cancer treatment. Previous anatomical and

clinical studies have demonstrated the usefulness of inguinal lymph node

autotransplantation in managing lymphoedema. The present anatomical study is a

complement to previous studies about the cutaneous inguinal flap. It has

demonstrated the feasibility of using a free inguinal cutaneo-lymph node flap

supplied by the superficial circumflex iliac artery. The useful vascularized

abdominal skin area ranged from 176 to 288 cm2 and was contained within a

vascularized skin area extending up to 928 cm2. However, the vascularization

never extended widely beyond the abdominal midline line. Although it mainly

remains unilateral, this flap combining skin and lymph nodes may help in the

management of lymphoedema and chest wall damage when they occur simultaneously as

complications of breast cancer treatment. The French version of this article is

available in the form of electronic supplementary material and can be obtained by

using the Springer LINK server located at

http://dx.doi.org/10.1007/s00276-002-0024-7.

PMID: 12197025 [Indexed for MEDLINE]

3584. Surg Radiol Anat. 2002 May;24(2):81-6.

Anatomical study of the blood supply of the coxal bone: radiological and clinical

application.

Yiming A(1), Baqué P, Rahili A, Mayer J, Braccini AL, Fontaine A, Leplatois A,

Clavé A, Bourgeon A, de Peretti F.

Author information:

(1)Laboratoire d'Anatomie, Faculté de Médecine, Avenue de Vallombrose, 06000

Nice, France.

The aim of this work was to study the arterial blood supply of the coxal bone in

order to optimize radiological embolization and to minimize the risk of

postoperative osteonecrosis. Ten fresh cadavers were dissected after

intra-arterial injection of colored resin. All the collateral vessels running to

this bone were described and counted. On 25 dry bones, the vascular foramina were

measured with the aid of a millimetric gauge and a vascular map was created. The

posterior part of the ilium appears to be twice as well vascularized as the

anterior part. Fractures of the posterior arch of the pelvis are theoretically

more hemorrhagic. The presence of the iliolumbar artery in contact with the

sacroiliac joint increases the risk with open book or shearing fractures. The

artery of the ischium, a collateral of the pudendal artery, supplies the

posterior and lateral parts of the acetabulum and the artery of the roof of the

acetabulum, its superior and lateral parts. The branches of the anterior and

posterior divisions of the obturator artery supply the superior part of the

surroundings of the obturator foramen and the antero-inferior and

postero-inferior parts of the acetabulum. The Kocher approach may injure the

artery of the ischium. Letournel's extended lateral approach and Mears'

triradiate approach may injure the artery of the ischium and the artery of the

roof of the acetabulum. The risk of osteonecrosis appears to be theoretically

increased if one adds an endopelvic approach. The anterior approach to the

acetabulum appears to be that which theoretically leads to the least

devascularization. The French version of this article is available in the form of

electronic supplementary material and can be obtained by using the Springer Link

server located at http://dx.doi.org/10.1007/s00276-002-0029-2.

PMID: 12197024 [Indexed for MEDLINE]

3585. Surg Radiol Anat. 2002 May;24(2):129-32.

High duplication of the internal jugular vein: clinical incidence in the adult

and surgical consequences, a report of three clinical cases.

Prades JM(1), Timoshenko A, Dumollard JM, Durand M, Merzougui N, Martin C.

Author information:

(1)Laboratoire d'Anatomie, UFR de Médecine Jacques Lisfranc, 15, rue Ambroise

Paré, 42023, Saint-Etienne, France. prades@univ-st-etienne.fr

Duplication of the internal jugular vein (IJV) is a rare malformation. Three

intraoperative cases are reported. In our personal experience, the clinical

incidence of the anomaly is approximately 4 per 1,000 unilateral neck

dissections. The venous duplication is at a variable height, affecting the

superior part of the IJV. The lateral branch of the accessory nerve (XI) always

passes medially to the anterior vein and laterally to the posterior vein, between

the venous duplication. This is most often unilateral but sometimes bilateral.

The IJV may be normal, dilated or ectatic. The discovery of this anatomical

variation has practical implications during cervical lymph node clearance, either

functional or radical, during oncological surgery necessitating viewing the IJV

and its affluents and the lateral branch of the accessory nerve. The

embryological explanation suggests a topographical "conflict" between the

development of the IJV and the lateral branch of the accessory nerve. The French

version of this article is available in the form of electronic supplementary

material and can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00276-002-0020-y.

PMID: 12197023 [Indexed for MEDLINE]

3586. Surg Radiol Anat. 2002 May;24(2):125-8.

Multiple arterial variation of the human upper limb.

Durgun B(1), Yücel AH, Kizilkanat ED, Dere F.

Author information:

(1)Department of Anatomy, Faculty of Medicine, Cukurova University, 01330 Adana,

Turkey. bdurgun@cu.edu.tr

A case report of multiple variations involving the arteries of the upper limb in

a single cadaver is presented. In addition to the proximal origin of the arteries

unusual arterial patterns on both the right and left sides were present. On the

right side, the subscapular artery gave rise to a large posterior circumflex

humeral artery in addition to the thoracodorsal and circumflex scapular arteries.

On the left side, the radial and ulnar arteries arose from the brachial artery at

the level of arm, with their origins being opposite to the usual arrangement.

There was an arciform anastomosis between the radial and ulnar arteries, with the

radial recurrent artery arising from the concavity of the arch. The course of

both the radial and ulnar arteries was normal at the wrist and hand, except for

the absence of the first palmar metacarpal artery and an early bifurcation of the

second palmar metacarpal artery. The French version of this article is available

in the form of electronic supplementary material and can be obtained by using the

Springer LINK server located at http://dx.doi.org/10.1007/s00276-002-0011.

PMID: 12197022 [Indexed for MEDLINE]

3587. Surg Radiol Anat. 2002 May;24(2):120-4.

An MRI study of the meniscofemoral and transverse ligaments of the knee.

Erbagci H(1), Yildirim H, Kizilkan N, Gümüsburun E.

Author information:

(1)Department of Anatomy, Faculty of Medicine, Gaziantep University, Turkey.

erbagcihulya@hotmail.com

Our aim was to assess the anatomic localization, dimensions and incidence of the

transverse and meniscofemoral ligaments, which can show anatomic variations or be

mistaken for some pathologic conditions. In 100 healthy subjects (52 female, 48

male) whose ages ranged from 12 to 84 years, sagittal and coronal magnetic

resonance images of the knee were obtained. There was at least one anterior or

posterior meniscofemoral ligament in 82 cases. The anterior meniscofemoral

ligament was present in 8 of the female and 4 of the male subjects. The posterior

meniscofemoral ligament was found in 20 female and 22 male subjects. Both the

anterior and posterior meniscofemoral ligaments were present in 15 female and 13

male subjects. The transverse ligament of knee was encountered in 19 female and

12 male subjects. In the females, average lengths of the anterior and posterior

meniscofemoral ligaments were 9.87 +/- 4.79 mm and 25.60 +/- 5.50 mm,

respectively. The corresponding values in the males were 11.11 +/- 2.57 mm and

28.80 +/- 5.49 mm, respectively. In the females, average width of the anterior

and posterior meniscofemoral ligaments were 2.45 +/- 1.02 mm and 2.30 +/- 1.15

mm, respectively. The corresponding values in the males were 2.52 +/- 0.87 mm and

2.30 +/- 1.15 mm, respectively. On MRI assessment, in order to differentiate

intra-articular lesions such as osteochondral and meniscal fragments or

pseudotear of the lateral meniscus from the normal ligamentous anatomy of knee,

the orientation and characteristic localization of the meniscofemoral ligaments

should be taken into account. The French version of this article is available in

the form of electronic supplementary material and can be obtained by using the

Springer LINK server located at http://dx.doi.org/10.1007/s00276-002-0023-8.

PMID: 12197021 [Indexed for MEDLINE]

3588. Surg Radiol Anat. 2002 May;24(2):113-6.

Nervous branch passing through an accessory canal in the sphenozygomatic suture:

the temporal branch of the zygomatic nerve.

Akita K(1), Shimokawa T, Tsunoda A, Sato T.

Author information:

(1)Unit of Functional Anatomy, Graduate School, Tokyo Medical and Dental

University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8519, Japan.

akita.fana@tmd.ac.jp

A nervous branch which passes through a small canal in the sphenozygomatic suture

is sometimes observed during dissection. To examine the origin, course and

distribution of this nervous branch, 42 head halves of 21 Japanese cadavers (11

males, 10 females) and 142 head halves of 71 human dry skulls were used. The

branch was observed in seven sides (16.7%); it originated from the communication

between the lacrimal nerve and the zygomaticotemporal branch of the zygomatic

nerve or from the trunk of the zygomatic nerve. In two head halves (4.8%), the

branch pierced the anterior part of the temporalis muscle during its course to

the skin of the anterior part of the temple. The small canal in the suture was

observed in 31 head halves (21.8%) of the dry skulls. Although this nervous

branch is inconstantly observed, it should be called the temporal branch of the

zygomatic nerve according to the constant positional relationship to the sphenoid

and zygomatic bones. According to its origin, course and distribution, this

nervous branch may be considered to be influential in zygomatic and retro-orbital

pain due to entrapment and tension from the temporalis muscle and/or the narrow

bony canal. The French version of this article is available in the form of

electronic supplementary material and can be obtained by using the Springer LINK

server located at http://dx.doi.org/10.1007/s00276-002-0027-4.

PMID: 12197019 [Indexed for MEDLINE]

3589. Surg Radiol Anat. 2002 May;24(2):108-12.

Study of the relationship between the suprascapular artery and the brachial

plexus.

Dargaud J(1), Galichon V, Dargaud Y, Quesnel T, Morin A.

Author information:

(1)Laboratoire d'Anatomie, Faculté de Médecine Lyon Grange-Blanche, 8, avenue

Rockefeller, 69373 Lyon, France. fredon@cismsun.univ-lyon1.fr

The literature concerning the relationship of the suprascapular artery with the

brachial plexus is quite limited. The aims of this study were to investigate the

different types of relationships between the suprascapular artery and the

brachial plexus trunks and to try to clarify their prevalence in the European

population. Our study involved the dissection of 100 upper extremities from

adults, all of European origin (55 women, 45 men). In the classic description,

the suprascapular artery passes in front of the brachial plexus (group A) in the

majority of cases, then behind the omohyoid before reaching the superior border

of the scapula. There are two other types of relationship with the brachial

plexus: the suprascapular artery can pass between the trunks of the brachial

plexus (group B) or the artery can pass behind the brachial plexus (group C). To

our knowledge there has been only one study on the subject, carried out by Kosugi

et al. Our results show a not inconsiderable number of variations in the

relationships between the suprascapular artery and the brachial plexus (29%).

Although no differences were noted between the sexes, a significant number of

individual variations were revealed by the comparative study of dissections

performed on both upper extremities of the same individual. Our results suggest

that new studies are required to complete these initial anatomical results

observed on a small sample of subjects of European origin. The French version of

this article is available in the form of electronic supplementary material and

can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00276-002-0026-5.

PMID: 12197018 [Indexed for MEDLINE]

3590. J Mol Biol. 2002 Apr 5;317(4):541-57.

N-terminal N-myristoylation of proteins: prediction of substrate proteins from

amino acid sequence.

Maurer-Stroh S(1), Eisenhaber B, Eisenhaber F.

Author information:

(1)Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, Vienna, A-1030,

Austria. stroh@nt.imp.univie.ac.at

Myristoylation by the myristoyl-CoA:protein N-myristoyltransferase (NMT) is an

important lipid anchor modification of eukaryotic and viral proteins. Automated

prediction of N-terminal N-myristoylation from the substrate protein sequence

alone is necessary for large-scale sequence annotation projects but it requires a

low rate of false positive hits in addition to a sufficient sensitivity. Our

previous analysis of substrate protein sequence variability, NMT sequences and 3D

structures has revealed motif properties in addition to the known PROSITE motif

that are utilized in a new predictor described here. The composite prediction

function (with separate ad hoc parameterization (a) for queries from non-fungal

eukaryotes and their viruses and (b) for sequences from fungal species) consists

of terms evaluating amino acid type preferences at sequences positions close to

the N terminus as well as terms penalizing deviations from the physical property

pattern of amino acid side-chains encoded in multi-residue correlation within the

motif sequence. The algorithm has been validated with a self-consistency and two

jack-knife tests for the learning set as well as with kinetic data for model

substrates. The sensitivity in recognizing documented NMT substrates is above 95

% for both taxon-specific versions. The corresponding rate of false positive

prediction (for sequences with an N-terminal glycine residue) is close to 0.5 %;

thus, the technique is applicable for large-scale automated sequence database

annotation. The predictor is available as public WWW-server with the URL

http://mendel.imp.univie.ac.at/myristate/. Additionally, we propose a version of

the predictor that identifies a number of proteolytic protein processing sites at

internal glycine residues and that evaluates possible N-terminal myristoylation

of the protein fragments.A scan of public protein databases revealed new

potential NMT targets for which the myristoyl modification may be of critical

importance for biological function. Among others, the list includes kinases,

phosphatases, proteasomal regulatory subunit 4, kinase interacting proteins

KIP1/KIP2, protozoan flagellar proteins, homologues of mitochondrial translocase

TOM40, of the neuronal calcium sensor NCS-1 and of the cytochrome c-type heme

lyase CCHL. Analyses of complete eukaryote genomes indicate that about 0.5 % of

all encoded proteins are apparent NMT substrates except for a higher fraction in

Arabidopsis thaliana ( approximately 0.8 %).

Copyright 2002 Elsevier Science Ltd.

DOI: 10.1006/jmbi.2002.5426

PMID: 11955008 [Indexed for MEDLINE]

3591. Dev Genes Evol. 2002 Apr;212(3):141-4. Epub 2002 Mar 13.

Mox homeobox expression in muscle lineage of the gastropod Haliotis asinina:

evidence for a conserved role in bilaterian myogenesis.

Hinman VF(1), Degnan BM.

Author information:

(1)Deparment of Zoology and Entomology, University of Queensland 4072, Australia.

Mox homeobox genes are expressed during early vertebrate somitogenesis. Here we

describe the expression of Has-Mox, a Mox gene from the gastropod Haliotis

asinina. Has-Moxis expressed in the trochophore larva in paraxial mesodermal

bands. During larval development, Has-Mox expression remains restricted to

mesodermal cells destined to form adult muscle in the foot. This restricted

expression of Has-Mox in Haliotis is similar to that observed for vertebrate Mox

genes, suggesting a conserved role in myogenesis in deuterostomes and

lophotrochozoans. In contrast, Mox is not expressed in muscle lineages in the

ecdysozoan representatives Caenorhabditis elegans or Drosophila; the C.

elegansgenome has lost Mox altogether. Electronic supplementary material to this

paper can be obtained by using the Springer Link server located at

http://dx.doi.org/10.1007/s00427-002-0223-6.

DOI: 10.1007/s00427-002-0223-6

PMID: 11976952 [Indexed for MEDLINE]

3592. Dev Genes Evol. 2002 Apr;212(3):134-40. Epub 2002 Mar 7.

A conserved F-box gene with unusual transcript localization.

Das T(1), Purkayastha-Mukherjee C, D'Angelo J, Weir M.

Author information:

(1)Department of Biology, Wesleyan University, Middletown, CT 06459, USA.

Drosophila partner of paired ( ppa), which encodes an F-box protein that targets

the Pax transcription factor Paired (Prd) for degradation, has the striking

property that its mRNA is expressed in a striped pattern with a characteristic

registration relative to the striped expression of prd in early embryos.

Localized expression of F-box genes was not expected because F-box proteins

generally have multiple substrates. We hypothesize that the patterned mRNA

expression of Drosophila and zebrafish ppa homologs may reflect constraints

resulting from the localized expression of their degradation substrates. To begin

to test this idea, we wished to determine whether patterned mRNA expression is

commonly observed among F-box genes, or whether it might be peculiar to ppa and

its homologs, or even specific to Drosophila. We examined embryonic expression of

all predicted F-box genes in Drosophila and found that mRNAs of 21 out of 23

predicted F-box genes are expressed uniformly in early Drosophila embryos,

whereas ppa and CG4911 mRNAs are patterned, CG4911 being expressed at the

positions of gastrulation folds. We also identified and tested expression of ppa

in zebrafish, which has two highly conserved homologs, ppaA and ppaB, and found

that both are expressed during embryogenesis and have enriched mRNA expression in

regions including the neural tube, the head, and the fin buds. Despite being

unusual in having localized transcripts, we found that the Drosophila and

zebrafish homologs interact with the expected Drosophila components of the

cellular degradation machinery - Skp1 (SkpA) and Rbx1 (Roc1a) - suggesting that

the Ppa proteins are indeed functional F-box proteins. We conclude that

patterning of ppa mRNA could reflect a constraint on ppa function that is not

common among F-box genes. Electronic supplementary material to this paper can be

obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00427-002-0222-7.

DOI: 10.1007/s00427-002-0222-7

PMID: 11976951 [Indexed for MEDLINE]

3593. Genome Res. 2002 Apr;12(4):656-64.

BLAT--the BLAST-like alignment tool.

Kent WJ(1).

Author information:

(1)Department of Biology and Center for Molecular Biology of RNA, University of

California-Santa Cruz, Santa Cruz, CA 95064, USA. kent@biology.ucsc.edu

Analyzing vertebrate genomes requires rapid mRNA/DNA and cross-species protein

alignments. A new tool, BLAT, is more accurate and 500 times faster than popular

existing tools for mRNA/DNA alignments and 50 times faster for protein alignments

at sensitivity settings typically used when comparing vertebrate sequences.

BLAT's speed stems from an index of all nonoverlapping K-mers in the genome. This

index fits inside the RAM of inexpensive computers, and need only be computed

once for each genome assembly. BLAT has several major stages. It uses the index

to find regions in the genome likely to be homologous to the query sequence. It

performs an alignment between homologous regions. It stitches together these

aligned regions (often exons) into larger alignments (typically genes). Finally,

BLAT revisits small internal exons possibly missed at the first stage and adjusts

large gap boundaries that have canonical splice sites where feasible. This paper

describes how BLAT was optimized. Effects on speed and sensitivity are explored

for various K-mer sizes, mismatch schemes, and number of required index matches.

BLAT is compared with other alignment programs on various test sets and then used

in several genome-wide applications. http://genome.ucsc.edu hosts a web-based

BLAT server for the human genome.

DOI: 10.1101/gr.229202. Article published online before March 2002

PMCID: PMC187518

PMID: 11932250 [Indexed for MEDLINE]

3594. J Biol Inorg Chem. 2002 Apr;7(4-5):551-9. Epub 2002 Feb 8.

Characterization of a metalloregulatory bismuth(III) site in Staphylococcus

aureus pI258 CadC repressor.

Busenlehner LS(1), Apuy JL, Giedroc DP.

Author information:

(1)Department of Biochemistry and Biophysics, Center for Advanced Biomolecular

Research, Texas A&M University, College Station, TX 77843-2128, USA.

Staphylococcus aureus pI258 CadC is a metal sensor protein that regulates the

expression of the cad operon which encodes metal ion resistance proteins involved

in the efficient efflux of Cd(II), Pb(II), Zn(II) and, according to one report,

Bi(III) ions. In this paper, direct evidence is presented that Bi(III) binds to

CadC and negatively regulates cad operator/promoter (O/P) binding. Optical

absorption spectroscopy reveals that dimeric CadC binds approximately 0.8 mol

equivalents of Bi(III) per CadC monomer to form a coordination complex

characterized by three S(-)-->Bi(III) ligand-to-metal charge transfer

transitions, with the longest wavelength absorption band centered at 415 nm

(epsilon(415)=4000 M(Bi)(-1) cm(-1)). UV-Vis absorption spectra of wild-type and

mutant Cys-->Gly (Ser) substitution CadC mutants compared to [Bi(DTT)(2)],

[Bi(GSH)(3)] and [Bi(NAC)](3) model complexes reveal that Cys7, Cys11, Cys60 and

Cys58 directly coordinate Bi(III) in a tetrathiolate coordination complex. The

apparent affinity derived from a Bi(III)-displacement optical titration with

Cd(II) is estimated to be K(Bi)< or =10(12) M(-1). Apo-CadC binds with high

affinity [ K(a)=1.1(+/-0.3)x10(9) M(-1); 0.40 M NaCl, pH 7.0, 25 degrees C] to a

5'-fluorescein-labeled cad O/P oligonucleotide,while the binding of one molar

equivalent of Bi(III) per CadC monomer (Bi(1)-CadC) reduces the affinity by

approximately 170-fold. Strikingly, Bi(III)-responsive negative regulation of cad

O/P binding is abrogated for Bi(1)-C60G CadC and severely disrupted in Bi(1)-C7G

CadC, whose relative affinity is reduced only 10-fold. The mechanism of

Bi(III)-responsive metalloregulation is discussed, based on the findings

presented here. Electronic supplementary material to this paper can be obtained

by using the Springer Link server located at

http://dx.doi.org/10.1007/s00775-001-0336-9.

DOI: 10.1007/s00775-001-0336-9

PMID: 11941514 [Indexed for MEDLINE]

3595. J Biol Inorg Chem. 2002 Apr;7(4-5):548-50. Epub 2002 Jan 30.

The level of 8-oxo-7,8-dihydro-2'-deoxyguanosine is positively correlated with

the size of the labile iron pool in human lymphocytes.

Gackowski D(1), Kruszewski M, Bartlomiejczyk T, Jawien A, Ciecierski M, Olinski

R.

Author information:

(1)Department of Clinical Biochemistry, The Ludwik Rydygier Medical University in

Bydgoszcz, ul. Karlowicza 24, 85-092 Bydgoszcz, Poland.

It appears that the labile iron pool (LIP, low molecular weight iron) presence in

cells can result in the production of reactive oxygen species (ROS). ROS may be

responsible for the formation of 8-oxo-7,8-dihydro-2'-deoxyguanosine (8-oxodGuo)

in cellular DNA. In the present study we report on the relationship between LIP

and the endogenous level of 8-oxodGuo in human lymphocytes. Good correlation has

been determined between LIP and the oxidatively modified nucleoside. This in turn

points out the possibility that under physiological condition there is the

availability of LIP for catalyzing Fenton-type reactions in close proximity to

cellular DNA. Electronic supplementary material to this paper can be obtained by

using the Springer Link server located at

http://dx.doi.org/10.1007/s00775-001-0335-x.

DOI: 10.1007/s00775-001-0335-x

PMID: 11941513 [Indexed for MEDLINE]

3596. J Biol Inorg Chem. 2002 Apr;7(4-5):539-47. Epub 2002 Jan 31.

Solution structure of cyanoferricytochrome c: ligand-controlled conformational

flexibility and electronic structure of the heme moiety.

Yao Y(1), Qian C, Ye K, Wang J, Bai Z, Tang W.

Author information:

(1)State Key Laboratory of Coordination Chemistry, Nanjing University, Nanjing

210093, P.R. China.

The solution structure of cyanoferricytochrome c has been determined using NMR

spectroscopy. As a result of including additional constraints derived from

pseudocontact shifts, a high-resolution NMR structure was obtained with high

accuracy. In order to study the conformational transition between the native

protein and its ligand adducts, the present structure was compared with the

solution structures of the wild-type cytochrome c and the imidazole-cytochrome c

complex. Like the solution structure of imidazole-cytochrome c, the heme crevice

is widened by the swinging out of residues 77-85 and a noticeable shift of the

50s helix. However, unlike imidazole, cyanide exerts less significant

perturbation on the conformation of the heme cavity, which is revealed by a more

compact residue package in the distal pocket. Furthermore, comparison of the

solution structure of CN-iso-1Met80Ala cytochrome c with the structure of

cyanoferricytochrome c indicated that the binding of cyanide has a different

impact on the distal cavity conformation in the two proteins. In addition, the

magnetic properties of the present system are discussed and a comprehensive study

of the electronic structure of ligand-cytochrome c complexes and the native

protein is also described. Electronic supplementary material to this paper can be

obtained by using the Springer Link server located at

http://dx.doi.org/10.1007/s00775-001-0334-y.

DOI: 10.1007/s00775-001-0334-y

PMID: 11941512 [Indexed for MEDLINE]

3597. J Biol Inorg Chem. 2002 Apr;7(4-5):526-32. Epub 2002 Feb 13.

Iron-sulfur cluster biosynthesis: characterization of Schizosaccharomyces pombe

Isa1.

Wu G(1), Mansy SS, Hemann C, Hille R, Surerus KK, Cowan JA.

Author information:

(1)Evans Laboratory of Chemistry, Ohio State University, 100 West 18th Avenue,

Columbus, OH 43210, USA.

Eukaryotic Isa1 is one of several mitochondrial proteins that have been

implicated in Fe-S cluster assembly paths in vivo. We report the first

biochemical characterization of an eukaryotic member of this family and discuss

this in the context of results from in vivo studies and studies of bacterial

homologues. Schizosaccharomyces pombe Isa1 is a multimeric protein carrying

[2Fe-2S](2+) clusters that have been characterized by Mössbauer and optical

spectroscopic studies. Complex formation with a redox-active ferredoxin has been

identified through crosslinking experiments and the coordination chemistry and

stability of the native clusters has been investigated through site-directed

mutagenesis and spectroscopic analysis. Electronic supplementary material to this

paper, containing Mössbauer and UV-visible spectra for mutant Isa1 proteins, can

be obtained by using the Springer Link server located at

http://dx.doi.org/10.1007/s00775-001-0330-2.

DOI: 10.1007/s00775-001-0330-2

PMID: 11941510 [Indexed for MEDLINE]

3598. J Biol Inorg Chem. 2002 Apr;7(4-5):514-25. Epub 2002 Jan 23.

Hybrid cluster proteins (HCPs) from Desulfovibrio desulfuricans ATCC 27774 and

Desulfovibrio vulgaris (Hildenborough): X-ray structures at 1.25 A resolution

using synchrotron radiation.

Macedo S(1), Mitchell EP, Romão CV, Cooper SJ, Coelho R, Liu MY, Xavier AV,

LeGall J, Bailey S, Garner DC, Hagen WR, Teixeira M, Carrondo MA, Lindley P.

Author information:

(1)Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av.

República, EAN, Apartado 127, 2781-901 Oeiras, Portugal.

The structures of the hybrid cluster proteins (HCPs) from the sulfate-reducing

bacteria Desulfovibrio desulfuricans (ATCC 27774) and Desulfovibrio vulgaris

(Hildenborough) have been elucidated at a resolution of 1.25 A using X-ray

synchrotron radiation techniques. In the case of the D. desulfuricans protein,

protein isolation, purification, crystallization and X-ray data collection were

carried out under strict anaerobic conditions, whereas for the D. vulgaris

protein the conditions were aerobic. However, both structures are essentially the

same, comprising three domains and two iron-sulfur centres. One of these centres

situated near the exterior of the molecules in domain 1 is a cubane [4Fe-4S]

cluster, whereas the other, located at the interface of the three domains,

contains the unusual four-iron cluster initially found in the D. vulgaris

protein. Details of the structures and the associated EPR spectroscopy of the D.

desulfuricans protein are reported herein. These structures show that the nature

of the hybrid cluster, containing both oxygen and sulfur bridges, is independent

of the presence of oxygen in the isolation and crystallization procedure and also

does not vary significantly with changes in the oxidation state. The structures

and amino acid sequences of the HCP are compared with the recently elucidated

structure of the catalytic subunit of a carbon monoxide dehydrogenase from

Carboxydothermus hydrogenoformans and related dehydrogenases. Electronic

supplementary material to this paper can be obtained by using the Springer Link

server located at http://dx.doi.org/10.1007/s00775-001-0326-y.

DOI: 10.1007/s00775-001-0326-y

PMID: 11941509 [Indexed for MEDLINE]

3599. J Biol Inorg Chem. 2002 Apr;7(4-5):500-13. Epub 2002 Feb 14.

The nickel enzyme methyl-coenzyme M reductase from methanogenic archaea: In vitro

induction of the nickel-based MCR-ox EPR signals from MCR-red2.

Mahlert F(1), Bauer C, Jaun B, Thauer RK, Duin EC.

Author information:

(1)Max-Planck-Institut für terrestrische Mikrobiologie and Laboratorium für

Mikrobiologie, Fachbereich Biologie, Philipps-Universität,

Karl-von-Frisch-Strasse, 35043 Marburg, Germany.

Methyl-coenzyme M reductase (MCR) is a nickel enzyme catalyzing the formation of

methane from methyl-coenzyme M and coenzyme B in all methanogenic archaea. The

active purified enzyme exhibits the axial EPR signal MCR-red1 and in the presence

of coenzyme M and coenzyme B the rhombic signal MCR-red2, both derived from

Ni(I). Two other EPR-detectable states of the enzyme have been observed in vivo

and in vitro designated MCR-ox1 and MCR-ox2 which have quite different nickel EPR

signals and which are inactive. Until now the MCR-ox1 and MCR-ox2 states could

only be induced in vivo. We report here that in vitro the MCR-red2 state is

converted into the MCR-ox1 state by the addition of polysulfide and into a

light-sensitive MCR-ox2 state by the addition of sulfite. In the presence of O(2)

the MCR-red2 state was converted into a novel third state designated MCR-ox3 and

exhibiting two EPR signals similar but not identical to MCR-ox1 and MCR-ox2. The

formation of the MCR-ox states was dependent on the presence of coenzyme B.

Investigations with the coenzyme B analogues S-methyl-coenzyme B and

desulfa-methyl-coenzyme B indicate that for the induction of the MCR-ox states

the thiol group of coenzyme B is probably not of importance. The results were

obtained with purified active methyl-coenzyme M reductase isoenzyme I from

Methanothermobacter marburgensis. They are discussed with respect to the nickel

oxidation states in MCR-ox1, MCR-ox2 and MCR-ox3 and to a possible presence of a

second redox active group in the active site. Electronic supplementary material

to this paper can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00775-001-0325-z.

DOI: 10.1007/s00775-001-0325-z

PMID: 11941508 [Indexed for MEDLINE]

3600. J Biol Inorg Chem. 2002 Apr;7(4-5):461-72. Epub 2002 Jan 11.

Solution structure of the unbound, oxidized Photosystem I subunit PsaC,

containing [4Fe-4S] clusters F(A) and F(B): a conformational change occurs upon

binding to photosystem I.

Antonkine ML(1), Liu G, Bentrop D, Bryant DA, Bertini I, Luchinat C, Golbeck JH,

Stehlik D.

Author information:

(1)Present address: Institut für Experimentalphysik, Freie Universität Berlin,

Arnimallee 14, 14195 Berlin, Germany.

This work presents the three-dimensional NMR solution structure of recombinant,

oxidized, unbound PsaC from Synechococcus sp. PCC 7002. Constraints are derived

from homo- and heteronuclear one-, two- and three-dimensional (1)H and (15)N NMR

data. Significant differences are outlined between the unbound PsaC structure

presented here and the available X-ray structure of bound PsaC as an integral

part of the whole cyanobacterial PS I complex. These differences mainly concern

the arrangement of the N- and C-termini with respect to the [4Fe-4S] core domain.

In the NMR solution structure of PsaC the C-terminal region assumes a disordered

helical conformation, and is clearly different from the extended coil

conformation, which is one of the structural elements required to anchor PsaC to

the PS I core heterodimer. In solution the N-terminus of PsaC is in contact with

the pre-C-terminal region but slides in between the latter and the iron-sulfur

core region of the protein. Together, these features result in a concerted

movement of the N-terminus and pre-C-terminal region away from the F(A) binding

site, accompanied by a bending of the N-terminus. In comparison, the same

terminal regions are positioned much closer to F(A) and take up an anti-parallel

beta-sheet arrangement in PsaC bound to PS I. The conformational changes between

bound and unbound PsaC correlate with the differences reported earlier for the

EPR spectra of reduced F(A) and F(B) in bound versus unbound PsaC. The observed

different structural features in solution are highly relevant for unraveling the

stepwise assembly process of the stromal PsaC, PsaD and PsaE subunits to the PS I

core heterodimer. Electronic supplementary material to this paper can be obtained

by using the Springer Link server located at

http://dx.doi.org/10.1007/s00775-001-0321-3.

DOI: 10.1007/s00775-001-0321-3

PMID: 11941504 [Indexed for MEDLINE]

3601. J Biol Inorg Chem. 2002 Apr;7(4-5):451-60. Epub 2002 Jan 8.

Stability and nickel binding properties of peptides designed as scaffolds for the

stabilization of Ni(II)-Fe(4)S(4) bridged assemblies.

Laplaza CE(1), Holm RH.

Author information:

(1)Department of Chemistry and Chemical Biology, Harvard University, Cambridge,

MA 02138, USA.

Helix-loop-helix peptides containing 63 residues (HC(4)H(2), HC(4)HC, HC(5)H),

designated by their sequence and content of histidyl (H) and cysteinyl (C)

residues, have been previously synthesized for the purpose of stabilizing certain

bridged metal sites in proteins. These peptides bind one Fe(4)S(4) cluster by

means of a ferredoxin tricysteinyl consensus sequence and an additional Cys

residue, and one Ni(II) atom (HC(4)H(2), HC(5)H) in predesigned binding sites. In

this investigation, the apopeptides and their Fe(4)S(4) derivatives are shown to

be relatively stable to unfolding by guanidine hydrochloride, indicating

stability of secondary structure. With this property demonstrated, Ni(II) binding

equilibria have been evaluated in the terms of site-specific (Scatchard model)

and stepwise (stoichiometric) binding constants. Two peptides were designed to

have preformed CysHis(3) (HC(4)H(2)) and Cys(2)His(2) (HC(5)H) binding sites. The

data indicate one strong binding site in each peptide with preferred binding

constants k(1)=4.4x10(5) M(-1)(HC(4)H(2)) and 2.7x10(5) M(-1)(HC(5)H). Based on

X-ray absorption spectroscopic data, these binding steps are associated with the

formation of the desired coordination units Ni(II)CysHis(3) and

Ni(II)Cys(2)His(2). For peptide HC(4)HC, k(1)=2.5x10(5) M(-1), but the binding

site could not be fully identified. Collective evidence from this and prior

investigations supports the presence of the bridged assemblies Ni(II)-(mu(2)-S x

Cys)-[Fe(4)S(4)], stabilized by a scaffolding effect in peptides HC(4)H(2) and

HC(5)H. The assembly Ni(II)-X-[Fe(4)S(4)] is the minimal structure of the

A-Cluster of carbon monoxide dehydrogenase adduced from spectroscopic evidence;

bridge X is currently unidentified. These results suggest that de novo designed

peptides may serve as scaffolds for the construction of native bridged sites in

proteins. Electronic supplementary material to this paper can be obtained by

using the Springer Link server located at

http://dx.doi.org/10.1007/s00775-001-0320-4.

DOI: 10.1007/s00775-001-0320-4

PMID: 11941503 [Indexed for MEDLINE]

3602. J Biol Inorg Chem. 2002 Apr;7(4-5):384-96. Epub 2001 Dec 19.

In vitro study of the insulin-mimetic behaviour of vanadium(IV, V) coordination

compounds.

Rehder D(1), Costa Pessoa J, Geraldes CF, Castro MC, Kabanos T, Kiss T, Meier B,

Micera G, Pettersson L, Rangel M, Salifoglou A, Turel I, Wang D.

Author information:

(1)Institute of Inorganic and Applied Chemistry, University of Hamburg, 20146

Hamburg, Germany. rehder@xray.chemie.uni-hamburg.de

Erratum in

J Biol Inorg Chem 2002 Jun;7(6):675.

A representative set of vanadium(IV and V) compounds in varying coordination

environments has been tested in the concentration range 1 to 10(-6) mM, using

transformed mice fibroblasts (cell line SV 3T3), with respect to their short-term

cell toxicity (up to 36 hours) and their ability to stimulate glucose uptake by

cells. These insulin-mimetic tests have also been carried out with

non-transformed human fibroblasts (cell line F26). The compounds under

investigation comprise established insulin-mimetic species such as vanadate

([H(2)VO(4)](-)), [VO(acetylacetonate)(2)], [VO(2)(dipicolinate)](-) and

[VO(maltolate)(2)], and new systems and coordination compounds containing OO, ON,

OS, NS and ONS donor atom sets. A vitality test assay, measuring the reduction

equivalents released in the mitochondrial respiratory chain by intracellular

glucose degradation, is introduced and the results are counter-checked with

(3)H-labelled glucose. Most compounds are toxic at the 1 mM concentration level,

and most compounds are essentially non-toxic and about as effective as or more

potent than insulin at concentrations of 0.01 mM and below. V(V) compounds tend

to be less toxic than V(IV)compounds, and complexes containing thio functional

ligands are somewhat more toxic than others. Generally, ON ligation is superior

in insulin-mimetic efficacy to OO or O/ NS coordination, irrespective of the

vanadium oxidation state. There is, however, no striking correlation between the

nature of the ligand systems and the insulin-mimetic potency in these cell

culture tests, encompassing 41 vanadium compounds, the results on 22 of which are

reported in detail here. The syntheses and characteristics of various new

compounds are provided together with selected speciation results. The crystal and

molecular structures of [[VO(naph-tris)](2)] [where naph-tris is the Schiff base

formed between o-hydroxynaphthaldehyde and tris(hydroxymethyl)amine] are

reported. Electronic supplementary material to this paper can be obtained by

using the Springer Link server located at

http://dx.doi.org/10.1007/s00775-001-0311-5.

DOI: 10.1007/s00775-001-0311-5

PMID: 11941496 [Indexed for MEDLINE]

3603. Bioinformatics. 2002 Mar;18(3):498-9.

BetaTPred: prediction of beta-TURNS in a protein using statistical algorithms.

Kaur H(1), Raghava GP.

Author information:

(1)Institute of Microbial Technology, Sector-39A, Chandigarh, India.

MOTIVATION: beta-turns play an important role from a structural and functional

point of view. beta-turns are the most common type of non-repetitive structures

in proteins and comprise on average, 25% of the residues. In the past numerous

methods have been developed to predict beta-turns in a protein. Most of these

prediction methods are based on statistical approaches. In order to utilize the

full potential of these methods, there is a need to develop a web server.

RESULTS: This paper describes a web server called BetaTPred, developed for

predicting beta-TURNS in a protein from its amino acid sequence. BetaTPred allows

the user to predict turns in a protein using existing statistical algorithms. It

also allows to predict different types of beta-TURNS e.g. type I, I', II, II',

VI, VIII and non-specific. This server assists the users in predicting the

consensus beta-TURNS in a protein.

AVAILABILITY: The server is accessible from

http://imtech.res.in/raghava/betatpred/

PMID: 11934756 [Indexed for MEDLINE]

3604. Bioinformatics. 2002 Mar;18(3):496-7.

CASA: a server for the critical assessment of protein sequence alignment

accuracy.

Kahsay RY(1), Wang G, Dongre N, Gao G, Dunbrack RL Jr.

Author information:

(1)Delaware Biotechnology Institute and Department of Electrical and Computer

Engineering, University of Delaware, 15 Innovation Way, Newark, DE 19711, USA.

SUMMARY: A public server for evaluating the accuracy of protein sequence

alignment methods is presented. CASA is an implementation of the alignment

accuracy benchmark presented by Sauder et al. (Proteins, 40, 6-22, 2000). The

benchmark currently contains 39321 pairwise protein structure alignments produced

with the CE program from SCOP domain definitions. The server produces graphical

and tabular comparisons of the accuracy of a user's input sequence alignments

with other commonly used programs, such as BLAST, PSI-BLAST, Clustal W, and

SAM-T99.

AVAILABILITY: The server is located at http://capb.dbi.udel.edu/casa.

PMID: 11934755 [Indexed for MEDLINE]

3605. Bioinformatics. 2002 Mar;18(3):492-3.

GRIMM: genome rearrangements web server.

Tesler G(1).

Author information:

(1)Department of Computer Science and Engineering, University of California, San

Diego, CA 92093-0114, USA.

SUMMARY: Genome Rearrangements In Man and Mouse (GRIMM) is a tool for analyzing

rearrangements of gene orders in pairs of unichromosomal and multichromosomal

genomes, with either signed or unsigned gene data. Although there are several

programs for analyzing rearrangements in unichromosomal genomes, this is the

first to analyze rearrangements in multichromosomal genomes. GRIMM also provides

a new algorithm for analyzing comparative maps for which gene directions are

unknown.

AVAILABILITY: A web server, with instructions and sample data, is available at

http://www-cse.ucsd.edu/groups/bioinformatics/GRIMM.

PMID: 11934753 [Indexed for MEDLINE]

3606. Hum Mutat. 2002 Mar;19(3):225-33.

HbVar: A relational database of human hemoglobin variants and thalassemia

mutations at the globin gene server.

Hardison RC(1), Chui DH, Giardine B, Riemer C, Patrinos GP, Anagnou N, Miller W,

Wajcman H.

Author information:

(1)Department of Biochemistry and Molecular Biology, Pennsylvania State

University, University Park, Pennsylvania, USA.

We have constructed a relational database of hemoglobin variants and thalassemia

mutations, called HbVar, which can be accessed on the web at

http://globin.cse.psu.edu. Extensive information is recorded for each variant and

mutation, including a description of the variant and associated pathology,

hematology, electrophoretic mobility, methods of isolation, stability

information, ethnic occurrence, structure studies, functional studies, and

references. The initial information was derived from books by Dr. Titus Huisman

and colleagues [Huisman et al., 1996, 1997, 1998]. The current database is

updated regularly with the addition of new data and corrections to previous data.

Queries can be formulated based on fields in the database. Tables of common

categories of variants, such as all those involving the alpha1-globin gene (HBA1)

or all those that result in high oxygen affinity, are maintained by automated

queries on the database. Users can formulate more precise queries, such as

identifying "all beta-globin variants associated with instability and found in

Scottish populations." This new database should be useful for clinical diagnosis

as well as in fundamental studies of hemoglobin biochemistry, globin gene

regulation, and human sequence variation at these loci.

Copyright 2002 Wiley-Liss, Inc.

DOI: 10.1002/humu.10044

PMID: 11857738 [Indexed for MEDLINE]

3607. J Am Med Inform Assoc. 2002 Mar-Apr;9(2):105-15.

Roundtable on bioterrorism detection: information system-based surveillance.

Lober WB(1), Karras BT, Wagner MM, Overhage JM, Davidson AJ, Fraser H, Trigg LJ,

Mandl KD, Espino JU, Tsui FC.

Author information:

(1)University of Washington, Seattle, Washington, USA. lober@u.washington.edu

Comment in

J Am Med Inform Assoc. 2002 Mar-Apr;9(2):202-3.

During the 2001 AMIA Annual Symposium, the Anesthesia, Critical Care, and

Emergency Medicine Working Group hosted the Roundtable on Bioterrorism Detection.

Sixty-four people attended the roundtable discussion, during which several

researchers discussed public health surveillance systems designed to enhance

early detection of bioterrorism events. These systems make secondary use of

existing clinical, laboratory, paramedical, and pharmacy data or facilitate

electronic case reporting by clinicians. This paper combines case reports of six

existing systems with discussion of some common techniques and approaches. The

purpose of the roundtable discussion was to foster communication among

researchers and promote progress by 1) sharing information about systems,

including origins, current capabilities, stages of deployment, and architectures;

2) sharing lessons learned during the development and implementation of systems;

and 3) exploring cooperation projects, including the sharing of software and

data. A mailing list server for these ongoing efforts may be found at

http://bt.cirg.washington.edu.

PMCID: PMC344564

PMID: 11861622 [Indexed for MEDLINE]

3608. J Chem Inf Comput Sci. 2002 Mar-Apr;42(2):405-7.

A versatile structural domain analysis server using profile weight matrices.

Schmidt S(1), Bork P, Dandekar T.

Author information:

(1)EMBL, Postfach 102209, D-69012 Heidelberg, Germany.

The WEB tool "AnDom" assigns to a given protein sequence all experimentally

determined structural domains contained within it, including multidomain and

large proteins. The server uses profile specific matrices from custom generated

multiple sequence alignments of all known SCOP domains (SCOP version 1.50).

Prediction time is short allowing numerous applications for structural genomics

including investigation of complex eucaryotic protein families. The WWW server is

at http://www.bork.embl-heidelberg.de/AnDom, and profiles can be downloaded at

ftp.bork.embl-heidelberg.de/pub/users/ schmidt/AnDom.

PMID: 11911710 [Indexed for MEDLINE]

3609. Trends Genet. 2002 Mar;18(3):158-62.

SHOT: a web server for the construction of genome phylogenies.

Korbel JO(1), Snel B, Huynen MA, Bork P.

Author information:

(1)EMBL, Meyerhofstrasse 1, 69117, Heidelberg, Germany. korbel@embl-heidelberg.de

With the increasing availability of genome sequences, new methods are being

proposed that exploit information from complete genomes to classify species in a

phylogeny. Here we present SHOT, a web server for the classification of genomes

on the basis of shared gene content or the conservation of gene order that

reflects the dominant, phylogenetic signal in these genomic properties. In

general, the genome trees are consistent with classical gene-based phylogenies,

although some interesting exceptions indicate massive horizontal gene transfer.

SHOT is a useful tool for analysing the tree of life from a genomic point of

view. It is available at http://www.Bork.EMBL-Heidelberg.de/SHOT.

PMID: 11858840 [Indexed for MEDLINE]

3610. Arch Microbiol. 2002 Feb;177(2):139-49. Epub 2001 Dec 6.

Diversity of ammonia monooxygenase operon in autotrophic ammonia-oxidizing

bacteria.

Norton JM(1), Alzerreca JJ, Suwa Y, Klotz MG.

Author information:

(1)Department of Plants, Soils and Biometeorology, Utah State University, Logan,

UT 84322-4820, USA. jennyn@cc.usu.edu

Autotrophic ammonia-oxidizing bacteria use the essential enzyme ammonia

monooxygenase (AMO) to transform ammonia to hydroxylamine. The amo operon

consists of at least three genes, amoC, amoA, and amoB; amoA encodes the subunit

containing the putative enzyme active site. The use of the amo genes as

functional markers for ammonia-oxidizing bacteria in environmental applications

requires knowledge of the diversity of the amo operon on several levels: (1) the

copy number of the operon in the genome, (2) the arrangement of the three genes

in an individual operon, and (3) the primary sequence of the individual genes. We

present a database of amo gene sequences for pure cultures of ammonia-oxidizing

bacteria representing both the beta- and the gamma-subdivision of Proteobacteria

in the following genera: Nitrosospira (6 strains), Nitrosomonas (5 strains) and

Nitrosococcus (2 strains). The amo operon was found in multiple (2-3) nearly

identical copies in the beta-subdivision representatives but in single copies in

the gamma-subdivision ammonia oxidizers. The analysis of the deduced amino acid

sequence revealed strong conservation for all three Amo peptides in both primary

and secondary structures. For the amoA gene within the beta-subdivision,

nucleotide identity values are approximately 85% within the Nitrosomonas or the

Nitrosospira groups, but approximately 75% when comparing between these groups.

Conserved regions in amoA and amoC were identified and used as primer sites for

PCR amplification of amo genes from pure cultures, enrichments and the soil

environment. The intergenic region between amoC and amoA is variable in length

and may be used to profile the community of ammonia-oxidizing bacteria in

environmental samples. Electronic supplementary material to this paper can be

obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00203-001-0369-z.

DOI: 10.1007/s00203-001-0369-z

PMID: 11807563 [Indexed for MEDLINE]

3611. Bioinformatics. 2002 Feb;18(2):368-73.

The EBI SRS server--recent developments.

Zdobnov EM(1), Lopez R, Apweiler R, Etzold T.

Author information:

(1)EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome

Campus, Hinxton, Cambridgeshire CB10 1SD, UK. Evugeni.Zdobnov@EBI.ac.uk

MOTIVATION: The current data explosion is intractable without advanced data

management systems. The numerous data sets become really useful when they are

interconnected under a uniform interface--representing the domain knowledge. The

SRS has become an integration system for both data retrieval and applications for

data analysis. It provides capabilities to search multiple databases by shared

attributes and to query across databases fast and efficiently.

RESULTS: Here we present recent developments at the EBI SRS server

(http://srs.ebi.ac.uk). The EBI SRS server contains today more than 130

biological databases and integrates more than 10 applications. It is a central

resource for molecular biology data as well as a reference server for the latest

developments in data integration. One of the latest additions to the EBI SRS

server is the InterPro database-Integrated Resource of Protein Domains and

Functional Sites. Distributed in XML format it became a turning point in low

level XML-SRS integration. We present InterProScan as an example of data analysis

applications, describe some advanced features of SRS6, and introduce the

SRSQuickSearch JavaScript interfaces to SRS.

PMID: 11847095 [Indexed for MEDLINE]

3612. Eur Radiol. 2002 Feb;12(2):485-90. Epub 2001 Jul 4.

Intranet and radiology: a critical appraisal of radiological applications of

Intranet technology.

Achenbach S(1).

Author information:

(1)Samara Academy, Katharinenkirchhof 1, 20457 Hamburg, Germany.

s.achenbach@samara-academy.net

The World-Wide Web (WWW or Web) is the service which led to the huge popularity

of the Internet by making it user friendly. Already in the early years of the

Web, this technology was also used to make internal information systems easier to

use and hospitals and departments set up "Intranets". An Intranet consists of a

Web Server which is installed within a local area network (LAN) and allows

information retrieval with a Web browser. This paper highlights the different

fields of Intranet applications in radiology and in the hospital and focuses on

systems for organisational issues as well as for patient data distribution. While

an Intranet can be a solution for many problems, not only in radiology, it is

accompanied by serious threats and common misunderstandings which are discussed.

DOI: 10.1007/s003300100969

PMID: 11870454 [Indexed for MEDLINE]

3613. J Mol Model. 2002 Feb;8(2):58-64.

Homology modeling reveals the structural background of the striking difference in

thermal stability between two related [NiFe]hydrogenases.

Szilágyi A(1), Kovács KL, Rákhely G, Závodszky P.

Author information:

(1)Institute of Enzymology, Biological Research Center, Hungarian Academy of

Sciences, Karolina út 29, H-1113 Budapest, Hungary. szia@enzim.hu

Hydrogenases are redox metalloenzymes in bacteria that catalyze the uptake or

production of molecular hydrogen. Two homologous nickel-iron hydrogenases, HupSL

and HydSL from the photosynthetic purple sulfur bacterium Thiocapsa

roseopersicina, differ substantially in their thermal stabilities despite the

high sequence similarity between them. The optimum temperature of HydSL activity

is estimated to be at least 50 degrees C higher than that of HupSL. In this work,

homology models of both proteins were constructed and analyzed for a number of

structural properties. The comparison of the models reveals that the higher

stability of HydSL can be attributed to increased inter-subunit electrostatic

interactions: the homology models reliably predict that HydSL contains at least

five more inter-subunit ion pairs than HupSL. The subunit interface of HydSL is

more polar than that of HupSL, and it contains a few extra inter-subunit hydrogen

bonds. A more optimized cavity system and amino acid replacements resulting in

increased conformational rigidity may also contribute to the higher stability of

HydSL. The results are in accord with the general observation that with

increasing temperature, the role of electrostatic interactions in protein

stability increases. Electronic supplementary material to this paper can be

obtained by using the Springer Link server located at

http://dx.doi.org/10.1007/s00894-001-0071-8.

DOI: 10.1007/s00894-001-0071-8

PMID: 12032599 [Indexed for MEDLINE]

3614. Mol Genet Genomics. 2002 Feb;266(6):942-50. Epub 2001 Dec 15.

A novel MADS-box gene subfamily with a sister-group relationship to class B

floral homeotic genes.

Becker A(1), Kaufmann K, Freialdenhoven A, Vincent C, Li MA, Saedler H, Theissen

G.

Author information:

(1)Department of Molecular Plant Genetics, Max-Planck-Institute for Breeding

Research, Köln, Germany.

Class B floral homeotic genes specify the identity of petals and stamens during

the development of angiosperm flowers. Recently, putative orthologs of these

genes have been identified in different gymnosperms. Together, these genes

constitute a clade, termed B genes. Here we report that diverse seed plants also

contain members of a hitherto unknown sister clade of the B genes, termed

B(sister) (B(s)) genes. We have isolated members of the B(s) clade from the

gymnosperm Gnetum gnemon, the monocotyledonous angiosperm Zea mays and the

eudicots Arabidopsis thaliana and Antirrhinum majus. In addition, MADS-box genes

from the basal angiosperm Asarum europaeum and the eudicot Petunia hybrida were

identified as B(s) genes. Comprehensive expression studies revealed that B(s)

genes are mainly transcribed in female reproductive organs (ovules and carpel

walls). This is in clear contrast to the B genes, which are predominantly

expressed in male reproductive organs (and in angiosperm petals). Our data

suggest that the B(s) genes played an important role during the evolution of the

reproductive structures in seed plants. The establishment of distinct B and B(s)

gene lineages after duplication of an ancestral gene may have accompanied the

evolution of male microsporophylls and female megasporophylls 400-300 million

years ago. During flower evolution, expression of B(s) genes diversified, but the

focus of expression remained in female reproductive organs. Our findings imply

that a clade of highly conserved close relatives of class B floral homeotic genes

has been completely overlooked until recently and awaits further evaluation of

its developmental and evolutionary importance. Electronic supplementary material

to this paper can be obtained by using the Springer Link server located at

http://dx.doi.org/10.1007/s00438-001-0615-8.

DOI: 10.1007/s00438-001-0615-8

PMID: 11862488 [Indexed for MEDLINE]

3615. Naturwissenschaften. 2002 Feb;89(2):47-56.

Bose-Einstein condensation in dilute atomic gases.

Arlt JJ(1), Bongs K, Sengstock K, Ertmer W.

Author information:

(1)Institut für Quantenoptik, Universität Hannover, Germany.

arlt@iqo.uni-hannover.de

Bose-Einstein condensation is one of the most curious and fascinating phenomena

in physics. It lies at the heart of such intriguing processes as superfluidity

and superconductivity. However, in most cases, only a small part of the sample is

Bose-condensed and strong interactions are present. A weakly interacting, pure

Bose-Einstein condensate (BEC) has therefore been called the "holy grail of

atomic physics". In 1995 this grail was found by producing almost pure BECs in

dilute atomic gases. We review the experimental development that led to the

realization of BEC in these systems and explain how BECs are now routinely

produced in about 25 laboratories worldwide. The tremendous experimental progress

of the past few years is outlined and a number of recent experiments show the

current status of the field. Electronic supplementary material to this paper can

be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00114-001-0277-8.

PMID: 12046620

3616. Protein Sci. 2002 Feb;11(2):233-44.

The CATH extended protein-family database: providing structural annotations for

genome sequences.

Pearl FM(1), Lee D, Bray JE, Buchan DW, Shepherd AJ, Orengo CA.

Author information:

(1)Department of Biochemistry and Molecular Biology, University College London,

University of London, London WC1E 6BT, UK. frances@biochem.ucl.ac.uk

An automatic sequence search and analysis protocol (DomainFinder) based on

PSI-BLAST and IMPALA, and using conservative thresholds, has been developed for

reliably integrating gene sequences from GenBank into their respective structural

families within the CATH domain database

(http://www.biochem.ucl.ac.uk/bsm/cath\_new). DomainFinder assigns a new gene

sequence to a CATH homologous superfamily provided that PSI-BLAST identifies a

clear relationship to at least one other Protein Data Bank sequence within that

superfamily. This has resulted in an expansion of the CATH protein family

database (CATH-PFDB v1.6) from 19,563 domain structures to 176,597 domain

sequences. A further 50,000 putative homologous relationships can be identified

using less stringent cut-offs and these relationships are maintained within

neighbour tables in the CATH Oracle database, pending further evidence of their

suggested evolutionary relationship. Analysis of the CATH-PFDB has shown that

only 15% of the sequence families are close enough to a known structure for

reliable homology modeling. IMPALA/PSI-BLAST profiles have been generated for

each of the sequence families in the expanded CATH-PFDB and a web server has been

provided so that new sequences may be scanned against the profile library and be

assigned to a structure and homologous superfamily.

DOI: 10.1110/ps.16802

PMCID: PMC2373435

PMID: 11790833 [Indexed for MEDLINE]

3617. Teratology. 2002 Feb;65(2):78-87.

Design and development of an Internet registry for congenital heart defects.

Mitri W(1), Sandridge AL, Subhani S, Greer W.

Author information:

(1)Biostatistics, Epidemiology and Scientific Computing Department, King Faisal

Specialist Hospital and Research Centre, Riyadh 11211, Saudi Arabia.

BACKGROUND: Congenital Heart Defects (CHD) are conditions that encompass more

than 50 diagnoses and are due to developmental abnormalities early in fetal life.

The King Faisal Specialist Hospital and Research Centre in the Kingdom of Saudi

Arabia treats approximately 100 new cases per month. We recently developed a new

CHD Registry that captures, stores and processes our data via the Internet.

METHODS: The Registry was developed using Hypertext Markup Language (HTML),

Microsoft Active Server Pages and Microsoft Structured Query Language (SQL).

RESULTS: Details of CHD cases are captured in a World Wide Web (WWW) Registry,

permitting any browser-enabled PC or Mac to participate fully in all registry

functions, including data-entry, viewing, editing, searching, reporting,

validating, charting, and exporting data subsets to statistics packages. It

includes "administrative" features and an active security system. The paper forms

have been designed to reflect the "look and feel" of the Web pages. Automatic

validation procedures are also included.

CONCLUSIONS: Our Registry has been in operation for 3 years. It serves 10 PCs and

contains more than 3,000 registered cases of CHD. It is the first CHD Registry to

be fully functional on the Internet. It is also the first dedicated CHD registry,

and the first to routinely report on the full spectrum of CHD diagnoses. The WWW

offers several logistical advantages to disease registries, especially those that

represent large regions. It also offers the possibility of sharing resources

between registries, facilitating the aggregation and analysis of disease data on

a world-wide scale. This is useful for rare diseases such as CHD (see

http://rc.kfshrc.edu.sa/chdr/demo/).

Copyright 2002 Wiley-Liss, Inc.

DOI: 10.1002/tera.10016

PMID: 11857509 [Indexed for MEDLINE]

3618. Anal Bioanal Chem. 2002 Jan;372(1):205-15. Epub 2001 Dec 12.

A laser-induced fluorescence dual-fiber optic array detector applied to the rapid

HPLC separation of polycyclic aromatic hydrocarbons.

Hart SJ(1), Hall GJ, Kenny JE.

Author information:

(1)Tufts University, Chemistry Department, Medford. MA 02155, USA.

shart@ccf.nrl.navy.mil

A multi-channel detection system utilizing fiber optics has been developed for

the laser-induced fluorescence (LIF) analysis of chromatographic eluents. It has

been applied to the detection of polycyclic aromatic hydrocarbons (PAH) in a

chromatographically overlapped standard mixture and to a complex soil sample

extract obtained during fieldwork. The instrument utilizes dual-fiber optic

arrays, one to deliver multiple excitation wavelengths (258-342 nm) generated by

a Raman shifter, and the other to collect fluorescence generated by the sample at

each excitation wavelength; the collected fluorescence is dispersed and detected

with a spectrograph/CCD combination. The resulting data were arranged into

excitation emission matrices (EEM) for visualization and data analysis. Rapid

characterization of PAH mixtures was achieved under isocratic chromatographic

conditions (1.5 mL min(-1) and 80% acetonitrile in water), with mid microg L(-1)

detection limits, in less than 4 minutes. The ability of the instrument to

identify co-eluting compounds was demonstrated by identifying and quantifying

analytes in the rapid analysis of a 17 component laboratory-prepared PAH mixture

and a soil extracted sample. Identification and quantification were accomplished

using rank annihilation factor analysis (RAFA) using pure component standards and

the EEMs of mixtures measured during the rapid high-performance liquid

chromatography (HPLC) method as the unknowns. The percentage errors of the

retention times (RTs) determined using RAFA compared to the known RTs measured

with a standard absorbance detector were between 0 and 11%. For the standard PAH

mixture, all 17 components were identified correctly and for the soil extracted

sample, all 8 analytes present were correctly identified with only one false

positive. Overall, the system achieved excellent qualitative performance with

semi-quantitative results in the concentration predictions of both the standard

mixture and the real-world sample. Electronic supplementary material to this

paper can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00216-001-1125-6.

DOI: 10.1007/s00216-001-1125-6

PMID: 11939196

3619. Bioinformatics. 2002 Jan;18(1):213-4.

Geno3D: automatic comparative molecular modelling of protein.

Combet C(1), Jambon M, Deléage G, Geourjon C.

Author information:

(1)Institut de Biologie et Chimie des Protéines, IBCP-CNRS UMR 5086, Pôle

Bioinformatique Lyonnais, Bioinformatics and structural NMR, 7, Passage du

Vercors, 69 367 Lyon Cedex 07, France.

Geno3D (http://geno3d-pbil.ibcp.fr) is an automatic web server for protein

molecular modelling. Starting with a query protein sequence, the server performs

the homology modelling in six successive steps: (i) identify homologous proteins

with known 3D structures by using PSI-BLAST; (ii) provide the user all potential

templates through a very convenient user interface for target selection; (iii)

perform the alignment of both query and subject sequences; (iv) extract

geometrical restraints (dihedral angles and distances) for corresponding atoms

between the query and the template; (v) perform the 3D construction of the

protein by using a distance geometry approach and (vi) finally send the results

by e-mail to the user.

PMID: 11836238 [Indexed for MEDLINE]

3620. In Silico Biol. 2002;2(3):195-205.

AGenDA: gene prediction by comparative sequence analysis.

Rinner O(1), Morgenstern B.

Author information:

(1)GSF Research Center, MIPS/Institute of Bioinformatics, Ingolstädter

Landstrasse 1, 85764 Neuherberg, Germany.

Comparative sequence analysis is a powerful approach to identify functional

elements in genomic sequences. Herein, we describe AGenDA (Alignment-based GENe

Detection Algorithm), a novel method for gene prediction that is based on

long-range alignment of syntenic regions in eukaryotic genome sequences. Local

sequence homologies identified by the DIALIGN program are searched for conserved

splice signals to define potential protein-coding exons; these candidate exons

are then used to assemble complete gene structures. The performance of our method

was tested on a set of 105 human-mouse sequence pairs. These test runs showed

that sensitivity and specificity of AGenDA are comparable with the best gene-

prediction program that is currently available. However, since our method is

based on a completely different type of input information, it can detect genes

that are not detectable by standard methods and vice versa. Thus, our approach

seems to be a useful addition to existing gene-prediction programs.AVAILABILITY:

DIALIGN is available through the Bielefeld Bioinformatics Server (BiBiServ) at

http://bibiserv.techfak.uni-bielefeld.de/dialign/ The gene-prediction program

AGenDA described in this paper will be available through the BiBiServ or MIPS web

server at http://mips.gsf.de.

PMID: 12542405 [Indexed for MEDLINE]

3621. J Digit Imaging. 2002;15 Suppl 1:246-9. Epub 2002 Mar 21.

OpenRIMS: an open architecture radiology informatics Management system.

Langer SG(1).

Author information:

(1)University of Washington Seattle, 1959 NE Pacific Street, 98195, USA.

sglanger@u.washington.edu

The following are benefits of an integrated picture archiving and communication

system/radiology information system archive built with open-source tools and

methods: open source, inexpensive interfaces can be updated as needed, and

reduced risk of redundant and inconsistent data. Also, wide adoption would

promote standard data mining tools, reducing user needs to learn multiple methods

to perform the same task. A model has been constructed capable of accepting

orders, performing exam resource scheduling, providing Digital communications in

Medicine (DICOM) work list information to modalities, archiving studies, and

supporting DICOM query/retrieve from third-party viewing software. The

multitiered architecture uses a single database communicating via an open

database connectivity bridge to a Linux server with Health Level 7 (HL7), DICOM,

and HTTP connections. Human interaction is supported via a browser, whereas other

informatics systems communicate over the HL7 and DICOM links. The system is still

under development, but the primary database schema is complete, as are key pieces

of the Web user interface. Additional work is needed on the DICOM/HL7 interface

broker and completion of the base DICOM service classes.

DOI: 10.1007/s10278-002-5008-y

PMID: 12105742 [Indexed for MEDLINE]

3622. J Digit Imaging. 2002;15 Suppl 1:201-5. Epub 2002 Mar 21.

The completeness of existing lexicons for representing radiology report

information.

Langlotz CP(1), Caldwell SA.

Author information:

(1)Department of Radiology, University of Pennsylvania, Philadelphia, USA.

langltoz@rad.upenn.edu

Although most medical lexicons contain up to 80% of clinical terms used in an

ambulatory patient medical records archive, preliminary research suggests that

they may be far less complete for radiology terms. We therefore compared the

likelihood that several existing medical lexicons would contain terms found in a

radiology report to the likelihood they would contain terms found in an

ambulatory care medical record. We used three samples of imaging terms to assess

the completeness of existing lexicons for medical imaging: (1) a random sample of

imaging terms from the Unified Medical Language System Large Scale Vocabulary

Test (UMLS-LSVT; n = 218), (2) terms actually used in the first 80 clinical knee

magnetic resonance imaging reports generated by the routine clinical use of a

structured reporting system (eDictation, Marlton, NJ; n = 76), and (3) terms

listed in a glossary of thoracic imaging prepared by the Fleischner Society (n =

173). Using the UMLS Web-based Knowledge Source Server (http://umlsks.nlm.nih.

gov/), we measured the rate at which terms in each of the above three sources

were found in the UMLS and two of its major constituent terminologies: ICD-9-CM

and SNOMED International. ICD-9-CM contained matches for 3%, 8%, and 11% of terms

from the Fleischner Society Glossary, eDictation, and NLM-LSVT, respectively.

SNOMED International contained matches for 32%, 46%, and 36% of terms from the

Fleischner Society Glossary, eDictation, and NLM-LSVT, respectively. The UMLS

contained matches for 36%, 50%, and 45% of terms from the Fleischner Society

Glossary, eDictation, and NLM-LSVT, respectively. The assessed vocabularies were

least likely to contain a term from the Fleischner Society Glossary and most

likely to contain a term from the eDictation lexicon. The UMLS was the most

complete, and ICD-9 was the least complete of the three systems evaluated. No

lexicon achieved greater than 50% completeness for any test set of imaging terms.

Our results show that no single lexicon is sufficiently complete to allow

comprehensive indexing, search, and retrieval of radiology report information.

These results confirm the few results available from the medical literature

indicating that existing controlled vocabularies are insufficiently complete to

represent the contents of radiology reports. A subjective analysis of these

results may identify particular imaging sub-areas for which new terms should be

developed.

DOI: 10.1007/s10278-002-5046-5

PMID: 12105728 [Indexed for MEDLINE]

3623. J Med Internet Res. 2002 Jan-Mar;4(1):e3.

Web-based cognitive behavior therapy: analysis of site usage and changes in

depression and anxiety scores.

Christensen H(1), Griffiths KM, Korten A.

Author information:

(1)The Centre for Mental Health Research, The Australian National University,

Canberra, ACT, Australia. Helen.Christensen@anu.edu.au

BACKGROUND: Cognitive behavior therapy is well recognized as an effective

treatment and prevention for depression when delivered face-to-face, via

self-help books (bibliotherapy), and through computer administration. The public

health impact of cognitive behavior therapy has been limited by cost and the lack

of trained practitioners. We have developed a free Internet-based cognitive

behavior therapy intervention (MoodGYM, http://moodgym.anu.edu.au) designed to

treat and prevent depression in young people, available to all Internet users,

and targeted to those who may have no formal contact with professional help

services.

OBJECTIVE: To document site usage, visitor characteristics, and changes in

depression and anxiety symptoms among users of MoodGYM, a Web site delivering a

cognitive-behavioral-based preventive intervention to the general public.

METHODS: All visitors to the MoodGYM site over about 6 months were investigated,

including 2909 registrants of whom 1503 had completed at least one online

assessment. Outcomes for 71 university students enrolled in an Abnormal

Psychology course who visited the site for educational training were included and

examined separately. The main outcome measures were (1) site-usage measures

including number of sessions, hits and average time on the server, and number of

page views; (2) visitor characteristics including age, gender, and initial

Goldberg self-report anxiety and depression scores; and (3) symptom change

measures based on Goldberg anxiety and depression scores recorded on up to 5

separate occasions.

RESULTS: Over the first almost-6-month period of operation, the server recorded

817284 hits and 17646 separate sessions. Approximately 20% of sessions lasted

more than 16 minutes. Registrants who completed at least one assessment reported

initial symptoms of depression and anxiety that exceeded those found in

population-based surveys and those characterizing a sample of University

students. For the Web-based population, both anxiety and depression scores

decreased significantly as individuals progressed through the modules.

CONCLUSIONS Web sites are a practical and promising means of delivering cognitive

behavioral interventions for preventing depression and anxiety to the general

public. However, randomized controlled trials are required to establish the

effectiveness of these interventions.

DOI: 10.2196/jmir.4.1.e3

PMCID: PMC1761927

PMID: 11956035 [Indexed for MEDLINE]

3624. J Mol Model. 2002 Jan;8(1):33-43.

Mesoscopic dynamics of colloids simulated with dissipative particle dynamics and

fluid particle model.

Dzwinel W(1), Yuen DA, Boryczko K.

Author information:

(1)AGH University of Mining and Metallurgy, Institute of Computer Science,

Kraków, Poland. dwitek@msi.umn.edu

We report results of numerical simulations of complex fluids, using a combination

of discrete-particle methods. Our molecular modeling repertoire comprises three

simulation techniques: molecular dynamics (MD), dissipative particle dynamics

(DPD), and the fluid particle model (FPM). This type of model can depict

multi-resolution molecular structures found in complex fluids ranging from single

micelle, colloidal crystals, large-scale colloidal aggregates up to the mesoscale

processes of hydrodynamical instabilities in the bulk of colloidal suspensions.

We can simulate different colloidal structures in which the colloidal beds are of

comparable size to the solvent particles. This undertaking is accomplished with a

two-level discrete particle model consisting of the MD paradigm with a

Lennard-Jones (L-J) type potential for defining the colloidal particle system and

DPD or FPM for modeling the solvent. We observe the spontaneous emergence of

spherical or rod-like micelles and their crystallization in stable hexagonal or

worm-like structures, respectively. The ordered arrays obtained by using the

particle model are similar to the 2D colloidal crystals observed in laboratory

experiments. The micelle shape and its hydrophobic or hydrophilic character

depend on the ratio between the scaling factors of the interactions between

colloid-colloid to colloid-solvent. Unlike the miscellar arrays, the colloidal

aggregates involve the colloid-solvent interactions prescribed by the DPD forces.

Different from the assumption of equilibrium growth, the two-level particle model

can display much more realistic molecular physics, which allows for the

simulation of aggregation for various types of colloids and solvent liquids over

a very broad range of conditions. We discuss the potential prospects of combining

MD, DPD, and FPM techniques in a single three-level model. Finally, we present

results from large-scale simulation of the Rayleigh-Taylor instability and

dispersion of colloidal slab in 2D and 3D. Electronic supplementary material to

this paper can be obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00894-001-0068-3.

DOI: 10.1007/s00894-001-0068-3

PMID: 12111400 [Indexed for MEDLINE]

3625. Naturwissenschaften. 2002 Jan;89(1):34-8.

Largest bird from the Early Cretaceous and its implications for the earliest

avian ecological diversification.

Zhou Z(1), Zhang F.

Author information:

(1)Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of

Sciences, Beijing. zhonghe@yeah.net

With only one known exception, early Cretaceous birds were smaller than their

closest theropod dinosaur relatives. Here we report on a new bird from the Early

Cretaceous feathered-dinosaur-bearing continental deposits of Liaoning, northeast

China, which is not only larger than Archaeopteryx but is nearly twice as large

as the basal dromaeosaur Microraptor. The new taxon, Sapeornis chaoyangensis gen.

et sp. nov., has a more basal phylogenetic position than all other birds except

for Archaeopteryx. Its exceptionally long forelimbs, well-developed deltoid crest

of the humerus, proximally fused metacarpals, relatively short hindlimbs and

short pygo-style indicate powerful soaring capability and further suggest that by

the Early Cretaceous ecological diversification of early birds was greater than

previously assumed. Electronic supplementary material to this paper can be

obtained by using the Springer LINK server located at

http://dx.doi.org/10.1007/s00114-001-0276-9.

PMID: 12008971 [Indexed for MEDLINE]

3626. Nucleic Acids Res. 2002 Jan 1;30(1):416-7.

RIDOM: Ribosomal Differentiation of Medical Micro-organisms Database.

Harmsen D(1), Rothgänger J, Frosch M, Albert J.

Author information:

(1)Institut für Hygiene und Mikrobiologie, Universität Würzburg, D-97080

Würzburg, Germany. dharmsen@ridom.de

The ribosomal differentiation of medical micro-organisms (RIDOM) web server,

first described by Harmsen et al. [Harmsden,D., Rothganger,J., Singer,C.,

Albert,J. and Frosch,M. (1999) Lancet, 353, 291], is an evolving electronic

resource designed to provide micro-organism differentiation services for medical

identification needs. The diagnostic procedure begins with a specimen partial

small subunit ribosomal DNA (16S rDNA) sequence. Resulting from a similarity

search, a species or genus name for the specimen in question will be returned.

Where the first results are ambiguous or do not define to species level, hints

for further molecular, i.e. internal transcribed spacer, and conventional

phenotypic differentiation will be offered ('sequential and polyphasic

approach'). Additionally, each entry in RIDOM contains detailed medical and

taxonomic information linked, context-sensitive, to external World Wide Web

services. Nearly all sequences are newly determined and the sequence

chromatograms are available for intersubjective quality control. Similarity

searches are now also possible by direct submission of trace files (ABI or SCF

format). Based on the PHRED/PHRAP software, error probability measures are

attached to each predicted nucleotide base and visualised with a new 'Trace

Editor'. The RIDOM web site is directly accessible on the World Wide Web at

http://www.ridom.de/. The email address for questions and comments is

webmaster@ridom.de.

PMCID: PMC99060

PMID: 11752353 [Indexed for MEDLINE]

3627. Nucleic Acids Res. 2002 Jan 1;30(1):383-4.

InBase: the Intein Database.

Perler FB(1).

Author information:

(1)New England Biolabs Inc., 32 Tozer Road, Beverly, MA 01915, USA.

perler@neb.com

Inteins are self-catalytic protein splicing elements. InBase

(http://www.neb.com/neb/inteins.html), the Intein Database and Registry, is a

curated compilation of published and unpublished information about protein

splicing. It presents general information as well as detailed data for each

intein, including tabulated comparisons and a comprehensive bibliography. An

intein-specific BLAST server is now available to assist in identifying new

inteins.

PMCID: PMC99080

PMID: 11752343 [Indexed for MEDLINE]

3628. Nucleic Acids Res. 2002 Jan 1;30(1):369-71.

AngioDB: database of angiogenesis and angiogenesis-related molecules.

Sohn TK(1), Moon EJ, Lee SK, Cho HG, Kim KW.

Author information:

(1)Graduate Program of Bioinformatics, Department of Molecular Biology, Pusan

National University, Pusan 609-735, Korea.

Angiogenesis is the formation of new capillaries sprouting from pre-existing

vessels. Angiogenesis occurs in a variety of normal physiological and

pathological conditions and is regulated by a balance of stimulatory and

inhibitory angiogenic factors. The control of this balance may fail and result in

the formation of a pathologic capillary network during the development of many

diseases. Therefore, we developed the angiogenesis database (AngioDB), which can

provide a signaling network of angiogenesis-related biomolecules in human. Each

record of AngioDB consisted of 12 fields and was developed by using a relational

database management system. For the retrieval of data, Active Server Page (ASP)

technology was integrated in this system. Users can access the database by a

query or imagemap browsing program. The retrieving system also provides a list of

angiogenesis-related molecules classified by three categories, and the database

has an external link to NCBI databases. AngioDB is available via the Internet at

http://angiodb.snu.ac.kr/.

PMCID: PMC99150

PMID: 11752339 [Indexed for MEDLINE]

3629. Nucleic Acids Res. 2002 Jan 1;30(1):364-8.

NUREBASE: database of nuclear hormone receptors.

Duarte J(1), Perrière G, Laudet V, Robinson-Rechavi M.

Author information:

(1)Laboratoire de Biologie Moléculaire et Cellulaire, CNRS UMR 5665, Ecole

Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France.

Nuclear hormone receptors are an abundant class of ligand activated

transcriptional regulators, found in varying numbers in all animals. Based on our

experience of managing the official nomenclature of nuclear receptors, we have

developed NUREBASE, a database containing protein and DNA sequences, reviewed

protein alignments and phylogenies, taxonomy and annotations for all nuclear

receptors. The reviewed NUREBASE is completed by NUREBASE\_DAILY, automatically

updated every 24 h. Both databases are organized under a client/server

architecture, with a client written in Java which runs on any platform. This

client, named FamFetch, integrates a graphical interface allowing selection of

families, and manipulation of phylogenies and alignments. NUREBASE sequence data

is also accessible through a World Wide Web server, allowing complex queries. All

information on accessing and installing NUREBASE may be found at

http://www.ens-lyon.fr/LBMC/laudet/nurebase.html.

PMCID: PMC99117

PMID: 11752338 [Indexed for MEDLINE]

3630. Nucleic Acids Res. 2002 Jan 1;30(1):294-8.

GTOP: a database of protein structures predicted from genome sequences.

Kawabata T(1), Fukuchi S, Homma K, Ota M, Araki J, Ito T, Ichiyoshi N, Nishikawa

K.

Author information:

(1)Center for Information Biology and DNA Data Bank of Japan, National Institute

of Genetics, 1-111 Yata, Mishima, Shizuoka 411-8540, Japan.

Large-scale genome projects generate an unprecedented number of protein

sequences, most of them are experimentally uncharacterized. Predicting the 3D

structures of sequences provides important clues as to their functions. We

constructed the Genomes TO Protein structures and functions (GTOP) database,

containing protein fold predictions of a huge number of sequences. Predictions

are mainly carried out with the homology search program PSI-BLAST, currently the

most popular among high-sensitivity profile search methods. GTOP also includes

the results of other analyses, e.g. homology and motif search, detection of

transmembrane helices and repetitive sequences. We have completed analyzing the

sequences of 41 organisms, with the number of proteins exceeding 120 000 in

total. GTOP uses a graphical viewer to present the analytical results of each ORF

in one page in a 'color-bar' format. The assigned 3D structures are presented by

Chime plug-in or RasMol. The binding sites of ligands are also included,

providing functional information. The GTOP server is available at

http://spock.genes.nig.ac.jp/~genome/gtop.html.

PMCID: PMC99104

PMID: 11752318 [Indexed for MEDLINE]

3631. Nucleic Acids Res. 2002 Jan 1;30(1):268-72.

SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence

searches, alignments and genome assignments.

Gough J(1), Chothia C.

Author information:

(1)MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

jgough@mrc-lmb.cam.ac.uk

The SUPERFAMILY database contains a library of hidden Markov models representing

all proteins of known structure. The database is based on the SCOP 'superfamily'

level of protein domain classification which groups together the most distantly

related proteins which have a common evolutionary ancestor. There is a public

server at http://supfam.org which provides three services: sequence searching,

multiple alignments to sequences of known structure, and structural assignments

to all complete genomes. Given an amino acid or nucleotide query sequence the

server will return the domain architecture and SCOP classification. The server

produces alignments of the query sequences with sequences of known structure, and

includes multiple alignments of genome and PDB sequences. The structural

assignments are carried out on all complete genomes (currently 59) covering

approximately half of the soluble protein domains. The assignments, superfamily

breakdown and statistics on them are available from the server. The database is

currently used by this group and others for genome annotation, structural

genomics, gene prediction and domain-based genomic studies.

PMCID: PMC99153

PMID: 11752312 [Indexed for MEDLINE]

3632. Nucleic Acids Res. 2002 Jan 1;30(1):255-9.

MODBASE, a database of annotated comparative protein structure models.

Pieper U(1), Eswar N, Stuart AC, Ilyin VA, Sali A.

Author information:

(1)Laboratories of Molecular Biophysics, The Pels Family Center for Biochemistry

and Structural Biology, The Rockefeller University, 1230 York Avenue, New York,

NY 10021, USA.

MODBASE (http://guitar.rockefeller.edu/modbase) is a relational database of

annotated comparative protein structure models for all available protein

sequences matched to at least one known protein structure. The models are

calculated by MODPIPE, an automated modeling pipeline that relies on PSI-BLAST,

IMPALA and MODELLER. MODBASE uses the MySQL relational database management system

for flexible and efficient querying, and the MODVIEW Netscape plugin for viewing

and manipulating multiple sequences and structures. It is updated regularly to

reflect the growth of the protein sequence and structure databases, as well as

improvements in the software for calculating the models. For ease of access,

MODBASE is organized into different datasets. The largest dataset contains models

for domains in 304 517 out of 539 171 unique protein sequences in the complete

TrEMBL database (23 March 2001); only models based on significant alignments

(PSI-BLAST E-value < 10(-4)) and models assessed to have the correct fold are

included. Other datasets include models for target selection and structure-based

annotation by the New York Structural Genomics Research Consortium, models for

prediction of genes in the Drosophila melanogaster genome, models for structure

determination of several ribosomal particles and models calculated by the MODWEB

comparative modeling web server.

PMCID: PMC99112

PMID: 11752309 [Indexed for MEDLINE]

3633. Nucleic Acids Res. 2002 Jan 1;30(1):31-4.

MIPS: a database for genomes and protein sequences.

Mewes HW(1), Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern

B, Münsterkötter M, Rudd S, Weil B.

Author information:

(1)Institute for Bioinformatics (MIPS), GSF National Research Center for

Environment and Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany.

w.mewes@gsf.de

The Munich Information Center for Protein Sequences (MIPS-GSF, Neuherberg,

Germany) continues to provide genome-related information in a systematic way.

MIPS supports both national and European sequencing and functional analysis

projects, develops and maintains automatically generated and manually annotated

genome-specific databases, develops systematic classification schemes for the

functional annotation of protein sequences, and provides tools for the

comprehensive analysis of protein sequences. This report updates the information

on the yeast genome (CYGD), the Neurospora crassa genome (MNCDB), the databases

for the comprehensive set of genomes (PEDANT genomes), the database of annotated

human EST clusters (HIB), the database of complete cDNAs from the DHGP (German

Human Genome Project), as well as the project specific databases for the GABI

(Genome Analysis in Plants) and HNB (Helmholtz-Netzwerk Bioinformatik) networks.

The Arabidospsis thaliana database (MATDB), the database of mitochondrial

proteins (MITOP) and our contribution to the PIR International Protein Sequence

Database have been described elsewhere [Schoof et al. (2002) Nucleic Acids Res.,

30, 91-93; Scharfe et al. (2000) Nucleic Acids Res., 28, 155-158; Barker et al.

(2001) Nucleic Acids Res., 29, 29-32]. All databases described, the protein

analysis tools provided and the detailed descriptions of our projects can be

accessed through the MIPS World Wide Web server (http://mips.gsf.de).

PMCID: PMC99165

PMID: 11752246 [Indexed for MEDLINE]

3634. J Mol Biol. 2001 Dec 14;314(5):1041-52.

Automatic clustering of orthologs and in-paralogs from pairwise species

comparisons.

Remm M(1), Storm CE, Sonnhammer EL.

Author information:

(1)Center for Genomics and Bioinformatics, Karolinska Institutet, S-17177

Stockholm, Sweden.

Orthologs are genes in different species that originate from a single gene in the

last common ancestor of these species. Such genes have often retained identical

biological roles in the present-day organisms. It is hence important to identify

orthologs for transferring functional information between genes in different

organisms with a high degree of reliability. For example, orthologs of human

proteins are often functionally characterized in model organisms. Unfortunately,

orthology analysis between human and e.g. invertebrates is often complex because

of large numbers of paralogs within protein families. Paralogs that predate the

species split, which we call out-paralogs, can easily be confused with true

orthologs. Paralogs that arose after the species split, which we call

in-paralogs, however, are bona fide orthologs by definition. Orthologs and

in-paralogs are typically detected with phylogenetic methods, but these are slow

and difficult to automate. Automatic clustering methods based on two-way best

genome-wide matches on the other hand, have so far not separated in-paralogs from

out-paralogs effectively. We present a fully automatic method for finding

orthologs and in-paralogs from two species. Ortholog clusters are seeded with a

two-way best pairwise match, after which an algorithm for adding in-paralogs is

applied. The method bypasses multiple alignments and phylogenetic trees, which

can be slow and error-prone steps in classical ortholog detection. Still, it

robustly detects complex orthologous relationships and assigns confidence values

for both orthologs and in-paralogs. The program, called INPARANOID, was tested on

all completely sequenced eukaryotic genomes. To assess the quality of INPARANOID

results, ortholog clusters were generated from a dataset of worm and mammalian

transmembrane proteins, and were compared to clusters derived by manual

tree-based ortholog detection methods. This study led to the identification with

a high degree of confidence of over a dozen novel worm-mammalian ortholog

assignments that were previously undetected because of shortcomings of

phylogenetic methods.A WWW server that allows searching for orthologs between

human and several fully sequenced genomes is installed at

http://www.cgb.ki.se/inparanoid/. This is the first comprehensive resource with

orthologs of all fully sequenced eukaryotic genomes. Programs and tables of

orthology assignments are available from the same location.

Copyright 2001 Academic Press.

DOI: 10.1006/jmbi.2000.5197

PMID: 11743721 [Indexed for MEDLINE]

3635. Bioinformatics. 2001 Dec;17(12):1242-3.

EVA: continuous automatic evaluation of protein structure prediction servers.

Eyrich VA(1), Martí-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F,

Valencia A, Sali A, Rost B.

Author information:

(1)Department of Chemistry, Columbia University, 3000 Broadway MC 3136, New York,

NY 10027, USA. eva@cubic.bioc.columbia.edu

Evaluation of protein structure prediction methods is difficult and

time-consuming. Here, we describe EVA, a web server for assessing protein

structure prediction methods, in an automated, continuous and large-scale

fashion. Currently, EVA evaluates the performance of a variety of prediction

methods available through the internet. Every week, the sequences of the latest

experimentally determined protein structures are sent to prediction servers,

results are collected, performance is evaluated, and a summary is published on

the web. EVA has so far collected data for more than 3000 protein chains. These

results may provide valuable insight to both developers and users of prediction

methods.AVAILABILITY: http://cubic.bioc.columbia.edu/eva.

CONTACT: eva@cubic.bioc.columbia.edu

PMID: 11751240 [Indexed for MEDLINE]

3636. Bioinformatics. 2001 Dec;17(12):1240-1.

ToolShop: prerelease inspections for protein structure prediction servers.

Rychlewski L(1).

Author information:

(1)Bioinformatics Laboratory, BioInfoBank Institute, ul. Limanowskiego 24A,

60-744 Poznan, Poland. leszek@bioinfo.pl

The ToolShop server offers a possibility to compare a protein tertiary structure

prediction server with other popular servers before releasing it to the public.

The comparison is conducted on a set of 203 proteins and the collected models are

compared with over 20 other programs using various assessment procedures. The

evaluation lasts circa one week.AVAILABILITY: The ToolShop server is available at

http://BioInfo.PL/ToolShop/. The administrator should be contacted to couple the

tested server to the evaluation suite.

CONTACT: leszek@bioinfo.pl

SUPPLEMENTARY INFORMATION: The evaluation procedures are similar to those

implemented in the continuous online server evaluation program, LiveBench.

Additional information is available from its homepage

(http://BioInfo.PL/LiveBench/).

PMID: 11751239 [Indexed for MEDLINE]

3637. Bioinformatics. 2001 Dec;17(12):1236-7.

ProPred: prediction of HLA-DR binding sites.

Singh H(1), Raghava GP.

Author information:

(1)Bioinformatics Centre, Institute of Microbial Technology, Sector 39A,

Chandigarh-160036, India.

ProPred is a graphical web tool for predicting MHC class II binding regions in

antigenic protein sequences. The server implement matrix based prediction

algorithm, employing amino-acid/position coefficient table deduced from

literature. The predicted binders can be visualized either as peaks in graphical

interface or as colored residues in HTML interface. This server might be a useful

tool in locating the promiscuous binding regions that can bind to several HLA-DR

alleles.AVAILABILITY: The server is available at

http://www.imtech.res.in/raghava/propred/

CONTACT: raghava@imtech.res.in

SUPPLEMENTARY INFORMATION: http://www.imtech.res.in/raghava/propred/page3.html

PMID: 11751237 [Indexed for MEDLINE]

3638. Bioinformatics. 2001 Dec;17(12):1226-7.

MELTING, computing the melting temperature of nucleic acid duplex.

Le Novère N(1).

Author information:

(1)Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2

3EJ, UK. nl223@cus.cam.ac.uk

MELTING computes the enthalpy and entropy of an oligonucleotide duplex helix-coil

transition, and then its melting temperature. The program uses the method of

nearest-neighbours. The set of thermodynamic parameters can be easily customized.

The program provides several correction methods for the concentration of salt.

MELTING is a free program, available at no cost and open-source. Perl scripts are

provided to show how MELTING can be used to construct more ambitious

programs.AVAILABILITY: MELTING is available for several platforms

(http://www.pasteur.fr/recherche/unites/neubiomol/meltinghome.html) and is

accessible via a www server

(http://bioweb.pasteur.fr/seqanal/interfaces/melting.html).

CONTACT: nl223@cus.cam.ac.uk

PMID: 11751232 [Indexed for MEDLINE]

3639. Dev Genes Evol. 2001 Dec;211(12):603-10. Epub 2001 Dec 15.

A zebrafish histone variant H2A.F/Z and a transgenic H2A.F/Z:GFP fusion protein

for in vivo studies of embryonic development.

Pauls S(1), Geldmacher-Voss B, Campos-Ortega JA.

Author information:

(1)Institut für Entwicklungsbiologie, Universität zu Köln, 50923 Cologne,

Germany.

Abstract. We have generated transgenic zebrafish lines expressing a fusion of a

histone variant, H2A.F/Z, to the green fluorescent protein (GFP) of the jellyfish

Aequorea victoria. Here, we describe the molecular cloning, partial

characterisation and expression of the zebrafish H2A.F/Z histone gene, as well as

the construction of the transgene and its transformation into the zebrafish germ

line. No abnormality can be detected in transgenic fish expressing the

H2A.F/Z:GFP fusion protein. The nuclear localisation of the fusion protein

correlates with the start of zygotic transcription, in that it is present in the

unfertilised egg and in the cytoplasm of cells after the first cleavages, being

found in some nuclei after the seventh or eighth cleavage, whereas all nuclei

from the 1,000-cell stage on, i.e. after midblastula transition, contain protein.

In addition to these data, we present a few examples of the many possible

applications of this transgenic line for developmental studies in vivo.

Electronic supplementary material to this paper can be obtained by using the

Springer LINK server located at http://dx.doi.org/10.1007/s00427-001-0196-x

DOI: 10.1007/s00427-001-0196-x

PMID: 11819118 [Indexed for MEDLINE]

3640. Protein Sci. 2001 Dec;10(12):2460-9.

Motif-based fold assignment.

Salwinski L(1), Eisenberg D.

Author information:

(1)Department of Chemistry, UCLA-DOE Laboratory of Structural Biology and

Molecular Medicine, UCLA, Los Angeles, California 90095-1570, USA.

Conventional fold recognition techniques rely mainly on the analysis of the

entire sequence of a protein. We present an MBA method to improve performance of

any conventional sequence-based fold assignment. The method uses sequence motifs,

such as those defined in the Prosite database, and the SwissProt annotation of

the fold library. When combined with a simple SDP method, the coverage of MBA is

comparable to the results obtained with PSI-BLAST. However, the set of the MBA

predictions is significantly different from that of PSI-BLAST, leading to a 40%

increase of the coverage for the combined MBA/PSI-BLAST method. The MBA approach

can be easily adopted to include the results of sequence-independent function

prediction methods and alternative motif and annotation databases. The method is

available through the web server localized at http://www.doe-mbi.ucla.edu/mba.

DOI: 10.1110/ps.14401

PMCID: PMC2374048

PMID: 11714913 [Indexed for MEDLINE]

3641. Proteins. 2001 Nov 15;45(3):241-61.

Linear programming optimization and a double statistical filter for protein

threading protocols.

Meller J(1), Elber R.

Author information:

(1)Department of Computer Science, Cornell University, Ithaca, New York 14853,

USA.

The design of scoring functions (or potentials) for threading, differentiating

native-like from non-native structures with a limited computational cost, is an

active field of research. We revisit two widely used families of threading

potentials: the pairwise and profile models. To design optimal scoring functions

we use linear programming (LP). The LP protocol makes it possible to measure the

difficulty of a particular training set in conjunction with a specific form of

the scoring function. Gapless threading demonstrates that pair potentials have

larger prediction capacity compared with profile energies. However, alignments

with gaps are easier to compute with profile potentials. We therefore search and

propose a new profile model with comparable prediction capacity to contact

potentials. A protocol to determine optimal energy parameters for gaps, using LP,

is also presented. A statistical test, based on a combination of local and global

Z-scores, is employed to filter out false-positives. Extensive tests of the new

protocol are presented. The new model provides an efficient alternative for

threading with pair energies, maintaining comparable accuracy. The code,

databases, and a prediction server are available at

http://www.tc.cornell.edu/CBIO/loopp.

Copyright 2001 Wiley-Liss, Inc.

PMID: 11599028 [Indexed for MEDLINE]

3642. J Mol Biol. 2001 Nov 2;313(4):903-19.

Assignment of homology to genome sequences using a library of hidden Markov

models that represent all proteins of known structure.

Gough J(1), Karplus K, Hughey R, Chothia C.

Author information:

(1)Laboratory of Molecular Biology, MRC, Hills Road, Cambridge, CB2 2QH, UK.

jgough@mrc-lmb.cam.ac.uk

Of the sequence comparison methods, profile-based methods perform with greater

selectively than those that use pairwise comparisons. Of the profile methods,

hidden Markov models (HMMs) are apparently the best. The first part of this paper

describes calculations that (i) improve the performance of HMMs and (ii)

determine a good procedure for creating HMMs for sequences of proteins of known

structure. For a family of related proteins, more homologues are detected using

multiple models built from diverse single seed sequences than from one model

built from a good alignment of those sequences. A new procedure is described for

detecting and correcting those errors that arise at the model-building stage of

the procedure. These two improvements greatly increase selectivity and coverage.

The second part of the paper describes the construction of a library of HMMs,

called SUPERFAMILY, that represent essentially all proteins of known structure.

The sequences of the domains in proteins of known structure, that have identities

less than 95 %, are used as seeds to build the models. Using the current data,

this gives a library with 4894 models. The third part of the paper describes the

use of the SUPERFAMILY model library to annotate the sequences of over 50

genomes. The models match twice as many target sequences as are matched by

pairwise sequence comparison methods. For each genome, close to half of the

sequences are matched in all or in part and, overall, the matches cover 35 % of

eukaryotic genomes and 45 % of bacterial genomes. On average roughly 15% of

genome sequences are labelled as being hypothetical yet homologous to proteins of

known structure. The annotations derived from these matches are available from a

public web server at: http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY. This server

also enables users to match their own sequences against the SUPERFAMILY model

library.

Copyright 2001 Academic Press.

DOI: 10.1006/jmbi.2001.5080

PMID: 11697912 [Indexed for MEDLINE]

3643. Bioinformatics. 2001 Nov;17(11):1077-83.

Database-driven multi locus sequence typing (MLST) of bacterial pathogens.

Chan MS(1), Maiden MC, Spratt BG.

Author information:

(1)The Wellcome Trust Centre for the Epidemiology of Infectious Disease,

University of Oxford, South Parks Road, Oxford OX1 3FY, UK.

mchan@molbiol.ox.ac.uk

MOTIVATION: Multi Locus Sequence Typing (MLST) is a newly developed typing method

for bacteria based on the sequence determination of internal fragments of seven

house-keeping genes. It has proved useful in characterizing and monitoring

disease-causing and antibiotic resistant lineages of bacteria. The strength of

this approach is that unlike data obtained using most other typing methods,

sequence data are unambiguous, can be held on a central database and be queried

through a web server.

RESULTS: A database-driven software system (mlstdb) has been developed, which is

used by public health laboratories and researchers globally to query their

nucleotide sequence data against centrally held databases over the internet. The

mlstdb system consists of a set of perl scripts for defining the database tables

and generating the database management interface and dynamic web pages for

querying the databases.

AVAILABILITY: http://www.mlst.net.

PMID: 11724739 [Indexed for MEDLINE]

3644. Protein Sci. 2001 Nov;10(11):2354-62.

Pcons: a neural-network-based consensus predictor that improves fold recognition.

Lundström J(1), Rychlewski L, Bujnicki J, Elofsson A.

Author information:

(1)Stockholm Bioinformatics Center, Stockholm University, SE 10691 Stockholm,

Sweden.

During recent years many protein fold recognition methods have been developed,

based on different algorithms and using various kinds of information. To examine

the performance of these methods several evaluation experiments have been

conducted. These include blind tests in CASP/CAFASP, large scale benchmarks, and

long-term, continuous assessment with newly solved protein structures. These

studies confirm the expectation that for different targets different methods

produce the best predictions, and the final prediction accuracy could be improved

if the available methods were combined in a perfect manner. In this article a

neural-network-based consensus predictor, Pcons, is presented that attempts this

task. Pcons attempts to select the best model out of those produced by six

prediction servers, each using different methods. Pcons translates the confidence

scores reported by each server into uniformly scaled values corresponding to the

expected accuracy of each model. The translated scores as well as the similarity

between models produced by different servers is used in the final selection.

According to the analysis based on two unrelated sets of newly solved proteins,

Pcons outperforms any single server by generating approximately 8%-10% more

correct predictions. Furthermore, the specificity of Pcons is significantly

higher than for any individual server. From analyzing different input data to

Pcons it can be shown that the improvement is mainly attributable to measurement

of the similarity between the different models. Pcons is freely accessible for

the academic community through the protein structure-prediction metaserver at

http://bioinfo.pl/meta/.

PMCID: PMC2374055

PMID: 11604541 [Indexed for MEDLINE]

3645. Bioinformatics. 2001 Sep;17(9):849-50.

The HMMTOP transmembrane topology prediction server.

Tusnády GE(1), Simon I.

Author information:

(1)Institute of Enzymology, BRC, Hungarian Academy of Sciences, H-1518 Budapest,

PO Box 7, Hungary. tusi@enzim.hu

The HMMTOP transmembrane topology prediction server predicts both the

localization of helical transmembrane segments and the topology of transmembrane

proteins. Recently, several improvements have been introduced to the original

method. Now, the user is allowed to submit additional information about segment

localization to enhance the prediction power. This option improves the prediction

accuracy as well as helps the interpretation of experimental results, i.e. in

epitope insertion experiments.AVAILABILITY: HMMTOP 2.0 is freely available to

non-commercial users at http://www.enzim.hu/hmmtop. Source code is also available

upon request to academic users.

PMID: 11590105 [Indexed for MEDLINE]

3646. Bioinformatics. 2001 Sep;17(9):847-8.

InterProScan--an integration platform for the signature-recognition methods in

InterPro.

Zdobnov EM(1), Apweiler R.

Author information:

(1)EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome

Campus, Hinxton, Cambridge CB10 1SD, UK. Evgueni.Zdobnov@EBI.ac.uk

InterProScan is a tool that scans given protein sequences against the protein

signatures of the InterPro member databases, currently--PROSITE, PRINTS, Pfam,

ProDom and SMART. The number of signature databases and their associated scanning

tools as well as the further refinement procedures make the problem complex.

InterProScan is designed to be a scalable and extensible system with a robust

internal architecture.AVAILABILITY: The Perl-based InterProScan implementation is

available from the EBI ftp server

(ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/) and the SRS-basedInterProScan is

available upon request. We provide the public web interface

(http://www.ebi.ac.uk/interpro/scan.html) as well as email submission server

(interproscan@ebi.ac.uk).

PMID: 11590104 [Indexed for MEDLINE]

3647. Bioinformatics. 2001 Sep;17(9):843-4.

GeneMachine: gene prediction and sequence annotation.

Makalowska I(1), Ryan JF, Baxevanis AD.

Author information:

(1)Genome Technology Branch, National Human Genome Research Institute, National

Institutes of Health, Building 49, Room 4A-22, Bethesda, MD 20892, USA.

MOTIVATION: A number of free-standing programs have been developed in order to

help researchers find potential coding regions and deduce gene structure for long

stretches of what is essentially 'anonymous DNA'. As these programs apply

inherently different criteria to the question of what is and is not a coding

region, multiple algorithms should be used in the course of positional cloning

and positional candidate projects to assure that all potential coding regions

within a previously-identified critical region are identified.

RESULTS: We have developed a gene identification tool called GeneMachine which

allows users to query multiple exon and gene prediction programs in an automated

fashion. BLAST searches are also performed in order to see whether a

previously-characterized coding region corresponds to a region in the query

sequence. A suite of Perl programs and modules are used to run MZEF, GENSCAN,

GRAIL 2, FGENES, RepeatMasker, Sputnik, and BLAST. The results of these runs are

then parsed and written into ASN.1 format. Output files can be opened using NCBI

Sequin, in essence using Sequin as both a workbench and as a graphical viewer.

The main feature of GeneMachine is that the process is fully automated; the user

is only required to launch GeneMachine and then open the resulting file with

Sequin. Annotations can then be made to these results prior to submission to

GenBank, thereby increasing the intrinsic value of these data.

AVAILABILITY: GeneMachine is freely-available for download at

http://genome.nhgri.nih.gov/genemachine. A public Web interface to the

GeneMachine server for academic and not-for-profit users is available at

http://genemachine.nhgri.nih.gov. The Web supplement to this paper may be found

at http://genome.nhgri.nih.gov/genemachine/supplement/.

PMID: 11590102 [Indexed for MEDLINE]

3648. Genome Res. 2001 Sep;11(9):1574-83.

SGP-1: prediction and validation of homologous genes based on sequence

alignments.

Wiehe T(1), Gebauer-Jung S, Mitchell-Olds T, Guigó R.

Author information:

(1)Max Planck Institute for Chemical Ecology, Jena, Germany. twiehe@ice.mpg.de

Conventional methods of gene prediction rely on the recognition of DNA-sequence

signals, the coding potential or the comparison of a genomic sequence with a

cDNA, EST, or protein database. Reasons for limited accuracy in many

circumstances are species-specific training and the incompleteness of reference

databases. Lately, comparative genome analysis has attracted increasing

attention. Several analysis tools that are based on human/mouse comparisons are

already available. Here, we present a program for the prediction of

protein-coding genes, termed SGP-1 (Syntenic Gene Prediction), which is based on

the similarity of homologous genomic sequences. In contrast to most existing

tools, the accuracy of depends little on species-specific properties such as

codon usage or the nucleotide distribution. may therefore be applied to

nonstandard model organisms in vertebrates as well as in plants, without the need

for extensive parameter training. In addition to predicting genes in large-scale

genomic sequences, the program may be useful to validate gene structure

annotations from databases. To this end, SGP-1 output also contains comparisons

between predicted and annotated gene structures in HTML format. The program can

be accessed via a Web server at http://soft.ice.mpg.de/sgp-1. The source code,

written in ANSI C, is available on request from the authors.

DOI: 10.1101/gr.177401

PMCID: PMC311140

PMID: 11544202 [Indexed for MEDLINE]

3649. Acta Crystallogr D Biol Crystallogr. 2001 Aug;57(Pt 8):1201-3. Epub 2001 Jul 23.

Molray--a web interface between O and the POV-Ray ray tracer.

Harris M(1), Jones TA.

Author information:

(1)Department of Cell and Molecular Biology, University of Uppsala, Box 590, 751

24 Uppsala, Sweden.

A publicly available web-based interface is presented for producing high-quality

ray-traced images and movies from the molecular-modelling program O [Jones et al.

(1991), Acta Cryst. A47, 110-119]. The interface allows the user to select O-plot

files and set parameters to create standard input files for the popular

ray-tracing renderer POV-Ray, which can then produce publication-quality still

images or simple movies. To ensure ease of use, we have made this service

available to the O user community via the World Wide Web. The public Molray

server is available at http://xray.bmc.uu.se/molray.

PMID: 11468417 [Indexed for MEDLINE]

3650. Bioinformatics. 2001 Aug;17(8):752-3.

Easier threading through web-based comparisons and cross-validations.

Douguet D(1), Labesse G.

Author information:

(1)Centre de Biochimie Structurale, INSERM U554 - CNRS UMR5048 - Université

Montpellier 1 15, Ave Charles Flahault, 34060 Montpellier Cedex, France.

douguet@cbs.univ-montp1.fr

We have developed a WWW server for the integration and comparison of protein

structure predictions performed by five different servers. Users submit an amino

acid sequence to a selected set of these prediction methods. Results are gathered

on a web-based page in order to facilitate comparison and analysis. All the

alignments are further evaluated through a common threading tool making their

comparisons easier.AVAILABILITY: The meta-server is available free at

http://www.infobiosud.cnrs.fr/bioserver

SUPPLEMENTARY INFORMATION: http://www.infobiosud.cnrs.fr/bioserver/hah1.html

PMID: 11524382 [Indexed for MEDLINE]

3651. Bioinformatics. 2001 Aug;17(8):750-1.

Structure prediction meta server.

Bujnicki JM(1), Elofsson A, Fischer D, Rychlewski L.

Author information:

(1)Bioinformatics Laboratory, International Institute of Molecular and Cell

Biology, Warsaw, Poland.

The Structure Prediction Meta Server offers a convenient way for biologists to

utilize various high quality structure prediction servers available worldwide.

The meta server translates the results obtained from remote services into uniform

format, which are consequently used to request a jury prediction from a remote

consensus server Pcons.AVAILABILITY: The structure prediction meta server is

freely available at http://BioInfo.PL/meta/, some remote servers have however

restrictions for non-academic users, which are respected by the meta server.

SUPPLEMENTARY INFORMATION: Results of several sessions of the CAFASP and

LiveBench programs for assessment of performance of fold-recognition servers

carried out via the meta server are available at http://BioInfo.PL/services.html.

PMID: 11524381 [Indexed for MEDLINE]

3652. Bioinformatics. 2001 Aug;17(8):748-9.

HOMSTRAD: adding sequence information to structure-based alignments of homologous

protein families.

de Bakker PI(1), Bateman A, Burke DF, Miguel RN, Mizuguchi K, Shi J, Shirai H,

Blundell TL.

Author information:

(1)Department of Biochemistry, University of Cambridge, 80 Tennis Court Road,

Cambridge CB2 1GA, UK.

summary: We describe an extension to the Homologous Structure Alignment Database

(HOMSTRAD; Mizuguchi et al., Protein Sci., 7, 2469-2471, 1998a) to include

homologous sequences derived from the protein families database Pfam (Bateman et

al., Nucleic Acids Res., 28, 263-266, 2000). HOMSTRAD is integrated with the

server FUGUE (Shi et al., submitted, 2001) for recognition and alignment of

homologues, benefitting from the combination of abundant sequence information and

accurate structure-based alignments. AVAILABILITY The HOMSTRAD database is

available at: http://www-cryst.bioc.cam.ac.uk/homstrad/. Query sequences can be

submitted to the homology recognition/alignment server FUGUE at:

http://www-cryst.bioc.cam.ac.uk/fugue/.

PMID: 11524380 [Indexed for MEDLINE]

3653. Bioinformatics. 2001 Aug;17(8):729-37.

The Bioinformatics Template Library--generic components for biocomputing.

Pitt WR(1), Williams MA, Steven M, Sweeney B, Bleasby AJ, Moss DS.

Author information:

(1)School of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX,

UK.

MOTIVATION: The efficiency of bioinformatics programmers can be greatly increased

through the provision of ready-made software components that can be rapidly

combined, with additional bespoke components where necessary, to create finished

programs. The new standard for C++ includes an efficient and easy to use library

of generic algorithms and data-structures, designed to facilitate low-level

component programming. The extension of this library to include functionality

that is specifically useful in compute-intensive tasks in bioinformatics and

molecular modelling could provide an effective standard for the design of

reusable software components within the biocomputing community.

RESULTS: A novel application of generic programming techniques in the form of a

library of C++ components called the Bioinformatics Template Library (BTL) is

presented. This library will facilitate the rapid development of efficient

programs by providing efficient code for many algorithms and data-structures that

are commonly used in biocomputing, in a generic form that allows them to be

flexibly combined with application specific object-oriented class libraries.

AVAILABILITY: The BTL is available free of charge from our web site

http://www.cryst.bbk.ac.uk/~classlib/ and the EMBL file server

http://www.embl-ebi.ac.uk/FTP/index.html

PMID: 11524374 [Indexed for MEDLINE]

3654. Bioinformatics. 2001 Aug;17(8):721-8.

Support vector machine approach for protein subcellular localization prediction.

Hua S(1), Sun Z.

Author information:

(1)Institute of Bioinformatics, State Key Laboratory of Biomembrane and Membrane

Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua

University, Beijing 100084, People's Republic of China.

huasj00@mails.tsinghua.edu.cn

MOTIVATION: Subcellular localization is a key functional characteristic of

proteins. A fully automatic and reliable prediction system for protein

subcellular localization is needed, especially for the analysis of large-scale

genome sequences.

RESULTS: In this paper, Support Vector Machine has been introduced to predict the

subcellular localization of proteins from their amino acid compositions. The

total prediction accuracies reach 91.4% for three subcellular locations in

prokaryotic organisms and 79.4% for four locations in eukaryotic organisms.

Predictions by our approach are robust to errors in the protein N-terminal

sequences. This new approach provides superior prediction performance compared

with existing algorithms based on amino acid composition and can be a

complementary method to other existing methods based on sorting signals.

AVAILABILITY: A web server implementing the prediction method is available at

http://www.bioinfo.tsinghua.edu.cn/SubLoc/.

SUPPLEMENTARY INFORMATION: Supplementary material is available at

http://www.bioinfo.tsinghua.edu.cn/SubLoc/.

PMID: 11524373 [Indexed for MEDLINE]

3655. Fresenius J Anal Chem. 2001 Aug;370(7):803-10.

web-based interactive data processing: application to stable isotope metrology.

Verkouteren RM(1), Lee JN.

Author information:

(1)National Institute of Standards and Technology, Gaithersburg MD 20899, USA.

To address a fundamental need in stable isotope metrology, the National Institute

of Standards and Technology (NIST) has established a web-based interactive

data-processing system accessible through a common gateway interface (CGI)

program on the internet site http://www. nist.gov/widps-co2. This is the first

application of a web-based tool that improves the measurement traceability

afforded by a series of NIST standard materials. Specifically, this tool promotes

the proper usage of isotope reference materials (RMs) and improves the quality of

reported data from extensive measurement networks. Through the International

Atomic Energy Agency (IAEA), we have defined standard procedures for stable

isotope measurement and data-processing, and have determined and applied

consistent reference values for selected NIST and IAEA isotope RMs. Measurement

data of samples and RMs are entered into specified fields on the web-based form.

These data are submitted through the CGI program on a NIST Web server, where

appropriate calculations are performed and results returned to the client.

Several international laboratories have independently verified the accuracy of

the procedures and algorithm for measurements of naturally occurring carbon-13

and oxygen-18 abundances and slightly enriched compositions up to approximately

150% relative to natural abundances. To conserve the use of the NIST RMs, users

may determine value assignments for a secondary standard to be used in routine

analysis. Users may also wish to validate proprietary algorithms embedded in

their laboratory instrumentation, or specify the values of fundamental variables

that are usually fixed in reduction algorithms to see the effect on the

calculations. The results returned from the web-based tool are limited in quality

only by the measurements themselves, and further value may be realized through

the normalization function. When combined with stringent measurement protocols,

two- to threefold improvements have been realized in the reproducibility of

carbon-13 and oxygen-18 determinations across laboratories.

PMID: 11569856

3656. Brain Res Mol Brain Res. 2001 Jul 13;91(1-2):169-73.

Radiation hybrid mapping of 11 alpha and beta nicotinic acetylcholine receptor

genes in Rattus norvegicus.

Tseng J(1), Kwitek-Black AE, Erbe CB, Popper P, Jacob HJ, Wackym PA.

Author information:

(1)Department of Otolaryngology and Communication Sciences, Medical College of

Wisconsin, 9200 W. Wisconsin Ave., Milwaukee, WI 53226, USA.

Acetylcholine is the main neurotransmitter of the vestibular efferents and a wide

variety of muscarinic and nicotinic acetylcholine receptors are expressed in the

vestibular periphery. To date, 11 nicotinic subunits (alpha and beta) have been

reported in mammals. Previously, our group [Brain Res. 778 (1997) 409] reported

that these nicotinic acetylcholine receptor alpha and beta subunits were

differentially expressed in the vestibular periphery of the rat. To begin an

understanding of the molecular genetics of these vestibular efferents, this study

examined the chromosomal locations of these nicotinic acetylcholine receptor

genes in the rat (Rattus norvegicus). Using radiation hybrid mapping and a rat

radiation hybrid map server (www.rgd.mcw.edu/RHMAP SERVER/), we determined the

chromosomal position for each of these genes. The alpha2-7, alpha9, alpha10, and

beta2-4 nicotinic subunits mapped to the following chromosomes: alpha2, chr. 15;

alpha3, chr. 8; alpha4, chr. 3; alpha5, chr. 8; alpha6, chr. 16; alpha7, chr. 1;

alpha9, chr. 14; alpha10, chr. 7; beta2, chr. 2; beta3, chr. 16; and beta4, chr.

8. With the location for each of these nicotinic subunits known, it is now

possible to develop consomic and/or congenic strains of rats that can be used to

study the functional genomics of each of these subunits.

PMID: 11457506 [Indexed for MEDLINE]

3657. Bioinformatics. 2001 Jul;17(7):656-7.

MEDUSA: large scale automatic selection and visual assessment of PCR primer

pairs.

Podowski RM(1), Sonnhammer EL.

Author information:

(1)Center for Genomics Research, Karolinska Institutet, S-171 77 Stockholm,

Sweden.

MEDUSA is a tool for automatic selection and visual assessment of PCR primer

pairs, developed to assist large scale gene expression analysis projects. The

system allows specification of constraints of the location and distances between

the primers in a pair. For instance, primers in coding, non-coding,

exon/intron-spanning regions might be selected. Medusa applies these constraints

as a filter to primers predicted by three external programs, and displays the

resulting primer pairs graphically in the Blixem (Sonnhammer and Durbin, COMPUT:

Appl. Biosci. 10, 301-307, 1994;

http://www.cgr.ki.se/cgr/groups/sonnhammer/Blixem.html) viewer.AVAILABILITY: The

MEDUSA web server is available at http://www.cgr.ki.se/cgr/MEDUSA. The source

code and user information are available at ftp://ftp.cgr.ki.se/pub/prog/medusa.

PMID: 11448885 [Indexed for MEDLINE]

3658. Bioinformatics. 2001 Jul;17(7):642-5.

Semi-automated update and cleanup of structural RNA alignment databases.

Gorodkin J(1), Zwieb C, Knudsen B.

Author information:

(1)Department of Ecology and Genetics, The Institute of Biological Sciences,

University of Aarhus, Building 540, Ny Munkegade, DK-8000 Aarhus C, Denmark.

We have developed a series of programs which assist in maintenance of structural

RNA databases. A main program BLASTs the RNA database against GenBank and

automatically extends and realigns the sequences to include the entire range of

the RNA query sequences. After manual update of the database, other programs can

examine base pair consistency and phylogenetic support. The output can be applied

iteratively to refine the structural alignment of the RNA database. Using these

tools, the number of potential misannotations per sequence was reduced from 20 to

3 in the Signal Recognition Particle RNA database.AVAILABILITY: A quick-server

and programs are available at http://www.bioinf.au.dk/rnadbtool/

PMID: 11448882 [Indexed for MEDLINE]

3659. Comput Biol Med. 2001 Jul;31(4):259-67.

ANTHEPROT: an integrated protein sequence analysis software with client/server

capabilities.

Deléage G(1), Combet C, Blanchet C, Geourjon C.

Author information:

(1)Pôle Bioinformatique Lyonnais, Institut de Biologie et Chimie des Protéines,

CNRS UMR 5086 Bioinformatique et RMN structurales, 7, Passage du Vercors, 69 367

Lyon, Cedex 07, France. g.deleage@ibcp.fr

Programs devoted to the analysis of protein sequences exist either as stand-alone

programs or as Web servers. However, stand-alone programs can hardly accommodate

for the analysis that involves comparisons on databanks, which require regular

updates. Moreover, Web servers cannot be as efficient as stand-alone programs

when dealing with real-time graphic display. We describe here a stand-alone

software program called ANTHEPROT, which is intended to perform protein sequence

analysis with a high integration level and clients/server capabilities. It is an

interactive program with a graphical user interface that allows handling of

protein sequence and data in a very interactive and convenient manner. It

provides many methods and tools, which are integrated into a graphical user

interface. ANTHEPROT is available for Windows-based systems. It is able to

connect to a Web server in order to perform large-scale sequence comparison on

up-to-date databanks. ANTHEPROT is freely available to academic users and may be

downloaded at http://pbil.ibcp.fr/ANTHEPROT.

PMID: 11334635 [Indexed for MEDLINE]

3660. Int J Med Inform. 2001 Jul;62(2-3):135-42.

Development and deployment of a web-based physician order entry system.

Lee YL(1), Hsu CY, Hsieh D, Li YC.

Author information:

(1)Taipei Medical College, Graduate Institute of Medical Informatics, Taipei,

Taiwan.

The computer-based Physician Order Entry System (POES) has been employed in many

clinical institutes in Taiwan. Most of the POES systems are developed in the

two-tier client-server architecture, and a large portion of the systems are

constructed on a mainframe or even a single PC. The exponential growth of the

Internet has had a tremendous impact on our society in recent years. In

consideration of the future user interface and system architecture, we have

developed a three-tier web-based Physician Order Entry System and successfully

deployed it in the Wang-Fang Hospital in Taipei. The system is the first POES

based on three-tier and World Wide Web (WWW) in Taiwan. The system provides the

Subjective, Objective, Assessment, and Plan (SOAP) structure for the physician to

enter subject, object, diagnoses, medicine dosage, treatment and laboratory test

request, and prints out the prescription and necessary document. The doctor can

also retrieve the patient's medical record on the system. One of the special

characteristics of the system is its personalized design. The doctor can define

their own diagnosis, medicine and treatment database and any combination of these

to facilitate their clinical work. The system has been reviewed since February

1999. The result shows that the clinical procedure has become more efficient, and

the chances of omission have been reduced. The system is very stable and the Open

Database Connectivity (ODBC) database access did not show any delay in the

network. Since we have incorporated many new web-programming techniques, the

progress of the techniques will improve the system performance in the future.

PMID: 11470616 [Indexed for MEDLINE]

3661. Protein Eng. 2001 Jul;14(7):465-72.

A Web-based classification system of DNA-binding protein families.

Karmirantzou M(1), Hamodrakas SJ.

Author information:

(1)Faculty of Biology, Department of Cell Biology and Biophysics, University of

Athens, Panepistimiopolis, Athens 157 01, Greece.

Rational classification of proteins encoded in sequenced genomes is critical for

making the genome sequences maximally useful for functional and evolutionary

studies. The family of DNA-binding proteins is one of the most populated and

studied amongst the various genomes of bacteria, archaea and eukaryotes and the

Web-based system presented here is an approach to their classification. The

DnaProt resource is an annotated and searchable collection of protein sequences

for the families of DNA-binding proteins. The database contains 3238 full-length

sequences (retrieved from the SWISS-PROT database, release 38) that include, at

least, a DNA-binding domain. Sequence entries are organized into families defined

by PROSITE patterns, PRINTS motifs and de novo excised signatures. Combining

global similarities and functional motifs into a single classification scheme,

DNA-binding proteins are classified into 33 unique classes, which helps to reveal

comprehensive family relationships. To maximize family information retrieval,

DnaProt contains a collection of multiple alignments for each DNA-binding family

while the recognized motifs can be used as diagnostically functional

fingerprints. All available structural class representatives have been

referenced. The resource was developed as a Web-based management system for

online free access of customized data sets. Entries are fully hyperlinked to

facilitate easy retrieval of the original records from the source databases while

functional and phylogenetic annotation will be applied to newly sequenced

genomes. The database is freely available for online search of a library

containing specific patterns of the identified DNA-binding protein classes and

retrieval of individual entries from our WWW server

(http://kronos.biol.uoa.gr/~mariak/dbDNA.html).

PMID: 11522919 [Indexed for MEDLINE]

3662. J Digit Imaging. 2001 Jun;14(2 Suppl 1):117-20.

"WWW.MDTF.ORG": a World Wide Web forum for developing open-architecture, freely

distributed, digital teaching file software by participant consensus.

Katzman GL(1), Morris D, Lauman J, Cochella C, Goede P, Harnsberger HR.

Author information:

(1)University of Utah Health Science Center, Center for Advanced Medical

Technologies, Salt Lake City 84108, USA.

PURPOSE: To foster a community supported evaluation processes for open-source

digital teaching file (DTF) development and maintenance. The mechanisms used to

support this process will include standard web browsers, web servers, forum

software, and custom additions to the forum software to potentially enable a

mediated voting protocol. The web server will also serve as a focal point for

beta and release software distribution, which is the desired end-goal of this

process.

CONCLUSIONS: We foresee that www.mdtf.org will provide for widespread

distribution of open source DTF software that will include function and interface

design decisions from community participation on the website forums.

PMCID: PMC3452713

PMID: 11442068 [Indexed for MEDLINE]

3663. Nucleic Acids Res. 2001 May 15;29(10):2135-44.

Discovering common stem-loop motifs in unaligned RNA sequences.

Gorodkin J(1), Stricklin SL, Stormo GD.

Author information:

(1)Department of Genetics and Ecology, The Institute of Biological Sciences,

University of Aarhus, Building 540, Ny Munkegade, DK-8000 Aarhus C, Denmark.

Post-transcriptional regulation of gene expression is often accomplished by

proteins binding to specific sequence motifs in mRNA molecules, to affect their

translation or stability. The motifs are often composed of a combination of

sequence and structural constraints such that the overall structure is preserved

even though much of the primary sequence is variable. While several methods exist

to discover transcriptional regulatory sites in the DNA sequences of coregulated

genes, the RNA motif discovery problem is much more difficult because of

covariation in the positions. We describe the combined use of two approaches for

RNA structure prediction, FOLDALIGN and COVE, that together can discover and

model stem-loop RNA motifs in unaligned sequences, such as UTRs from

post-transcriptionally coregulated genes. We evaluate the method on two datasets,

one a section of rRNA genes with randomly truncated ends so that a global

alignment is not possible, and the other a hyper-variable collection of IRE-like

elements that were inserted into randomized UTR sequences. In both cases the

combined method identified the motifs correctly, and in the rRNA example we show

that it is capable of determining the structure, which includes bulge and

internal loops as well as a variable length hairpin loop. Those automated results

are quantitatively evaluated and found to agree closely with structures contained

in curated databases, with correlation coefficients up to 0.9. A basic server,

Stem-Loop Align SearcH (SLASH), which will perform stem-loop searches in

unaligned RNA sequences, is available at http://www.bioinf.au.dk/slash/.

PMCID: PMC55461

PMID: 11353083 [Indexed for MEDLINE]

3664. Ir J Med Sci. 2001 Apr-Jun;170(2):123-5.

Designing a medical Web site.

Grannell MS(1), Singh RR, Tang R, Mansoor S, Walsh TN.

Author information:

(1)Department of Surgery, James Connolly Memorial Hospital, Blanchardstown,

Dublin, Ireland. markgrannell@eircom.net

BACKGROUND: A web site is a valuable shop window for any medical unit with

something to sell or something to say.

AIMS: The aim of this report is to outline the basic steps of web page design for

the individual or unit with limited financial resources.

METHODS: There are two ways of designing a web site. A reputable web design

company can be employed, but this is usually expensive. Alternatively, a web site

can be designed in-house using commercial software, following a few simple steps.

The basic requirements are a personal computer, software that is available on

most computers and access to a few peripheral items of hardware. An outline of

the page design should first be put down on paper. This can be transferred to a

computer file using a web page design program. This file is then sent to a server

for publication on the World Wide Web (WWW).

CONCLUSION: Designing and publishing a web page can take time and effort, but the

rewards can be great and the results will reflect the message and motto of the

unit.

PMID: 11491048 [Indexed for MEDLINE]

3665. J Pept Res. 2001 Apr;57(4):292-300.

Analysis of gammabeta, betagamma, gammagamma, betabeta continuous turns in

proteins.

Guruprasad K(1), Prasad MS, Kumar GR.

Author information:

(1)Center for DNA Fingerprinting and Diagnostics (CDFD), Nacharam, Hyderabad -

500 076, India. guru@gvk.com

We report the observation of continuous turns in proteins which comprise

individual gamma-turns or beta-turns or both that are situated immediately one

after the other along the polypeptide chain. The continuous turns were identified

from a representative data set of three-dimensional protein crystal structures.

The gammabeta/betagamma, gammagamma and betabeta continuous turns represent

peptides of varying amino acid residue lengths and conformations. The continuous

turns frequently observed in proteins were: gammabeta, between a coil and a

strand; betagamma, between a helix and a strand; gammagamma, between coils; and

betabeta, either between a strand and a coil or between strands or coils. We

determined the statistically significant amino acid residue preferences at

individual positions in the turn, calculated amino acid positional potentials and

analyzed main chain hydrogen bonds and side-chain interactions likely to

stabilize the continuous turns. The data on continuous turns have been integrated

in the database of structural motifs in proteins (DSMP) on our web server at

(http://www.cdfd.org.in/dsmp.html). This is useful to make queries on sequences

compatible with different continuous turns.

PMID: 11328486 [Indexed for MEDLINE]

3666. Nucleic Acids Res. 2001 Mar 15;29(6):1272-7.

Gene2EST: a BLAST2 server for searching expressed sequence tag (EST) databases

with eukaryotic gene-sized queries.

Gemünd C(1), Ramu C, Altenberg-Greulich B, Gibson TJ.

Author information:

(1)European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg,

Germany.

Expressed sequence tags (ESTs) are randomly sequenced cDNA clones. Currently,

nearly 3 million human and 2 million mouse ESTs provide valuable resources that

enable researchers to investigate the products of gene expression. The EST

databases have proven to be useful tools for detecting homologous genes, for exon

mapping, revealing differential splicing, etc. With the increasing availability

of large amounts of poorly characterised eukaryotic (notably human) genomic

sequence, ESTs have now become a vital tool for gene identification, sometimes

yielding the only unambiguous evidence for the existence of a gene expression

product. However, BLAST-based Web servers available to the general user have not

kept pace with these developments and do not provide appropriate tools for

querying EST databases with large highly spliced genes, often spanning 50 000-100

000 bases or more. Here we describe Gene2EST

(http://woody.embl-heidelberg.de/gene2est/), a server that brings together a set

of tools enabling efficient retrieval of ESTs matching large DNA queries and

their subsequent analysis. RepeatMasker is used to mask dispersed repetitive

sequences (such as Alu elements) in the query, BLAST2 for searching EST databases

and Artemis for graphical display of the findings. Gene2EST combines these

components into a Web resource targeted at the researcher who wishes to study one

or a few genes to a high level of detail.

PMCID: PMC29756

PMID: 11238992 [Indexed for MEDLINE]

3667. Protein Sci. 2001 Mar;10(3):599-612.

CODA: a combined algorithm for predicting the structurally variable regions of

protein models.

Deane CM(1), Blundell TL.

Author information:

(1)Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, United

Kingdom.

CODA, an algorithm for predicting the variable regions in proteins, combines

FREAD a knowledge based approach, and PETRA, which constructs the region ab

initio. FREAD selects from a database of protein structure fragments with

environmentally constrained substitution tables and other rule-based filters.

FREAD was parameterized and tested on over 3000 loops. The average root mean

square deviation ranged from 0.78 A for three residue loops to 3.5 A for eight

residue loops on a nonhomologous test set. CODA clusters the predictions from the

two independent programs and makes a consensus prediction that must pass a set of

rule-based filters. CODA was parameterized and tested on two unrelated separate

sets of structures that were nonhomologous to one another and those found in the

FREAD database. The average root mean square deviation in the test set ranged

from 0.76 A for three residue loops to 3.09 A for eight residue loops. CODA shows

a general improvement in loop prediction over PETRA and FREAD individually. The

improvement is far more marked for lengths six and upward, probably as the

predictive power of PETRA becomes more important. CODA was further tested on

several model structures to determine its applicability to the modeling

situation. A web server of CODA is available at

http://www-cryst.bioc.cam.ac.uk/~charlotte/Coda/search\_coda.html.

DOI: 10.1110/ps.37601

PMCID: PMC2374131

PMID: 11344328 [Indexed for MEDLINE]

3668. Proteins. 2001 Feb 15;42(3):319-31.

Defrosting the frozen approximation: PROSPECTOR--a new approach to threading.

Skolnick J(1), Kihara D.

Author information:

(1)Laboratory of Computational Genomics, Danforth Plant Science Center, Creve

Coeur, Missouri 63141, USA. skolnick@danfortcenter.org

PROSPECTOR (PROtein Structure Predictor Employing Combined Threading to Optimize

Results) is a new threading approach that uses sequence profiles to generate an

initial probe-template alignment and then uses this "partly thawed" alignment in

the evaluation of pair interactions. Two types of sequence profiles are used: the

close set, composed of sequences in which sequence identity lies between 35% and

90%; and the distant set, composed of sequences with a FASTA E-score less than

10. Thus, a total of four scoring functions are used in a hierarchical method:

the close (distant) sequence profiles screen a structural database to provide an

initial alignment of the probe sequence in each of the templates. The same

database is then screened with a scoring function composed of sequence plus

secondary structure plus pair interaction profiles. This combined hierarchical

threading method is called PROSPECTOR1. For the original Fischer database, 59 of

68 pairs are correctly identified in the top position. Next, the set of the top

20 scoring sequences (four scoring functions times the top five structures) is

used to construct a protein-specific pair potential based on consensus side-chain

contacts occurring in 25% of the structures. In subsequent threading iterations,

this protein-specific pair potential, when combined in a composite manner, is

found to be more sensitive in identifying the correct pairs than when the

original statistical potential is used, and it increases the number of recognized

structures for the combined scoring functions, termed PROSPECTOR2, to a total of

61 Fischer pairs identified in the top position. Application to a second, smaller

Fischer database of 27 probe-template pairs places 18 (17) structures in the top

position for PROSPECTOR1 (PROSPECTOR2). Overall, these studies show that the use

of pair interactions as assessed by the improved Z-score enhances the specificity

of probe-template matches. Thus, when the hierarchy of scoring functions is

combined, the ability to identify correct probe-template pairs is significantly

enhanced. Finally, a web server has been established for use by the academic

community (http://bioinformatics.danforthcenter.org/services/threading.html).

PMID: 11151004 [Indexed for MEDLINE]

3669. Bioinformatics. 2001 Feb;17(2):202-4.

RCNPRED: prediction of the residue co-ordination numbers in proteins.

Fariselli P(1), Casadio R.

Author information:

(1)CIRB Biocomputing Unit, Department of Biology, University of Bologna, via

Irnerio 42, 40126 Bologna, Italy. piero@lipid.biocomp.unibo.it

The RCNPRED server implements a neural network-based method to predict the

co-ordination numbers of residues starting from the protein sequence. Using

evolutionary information as input, RCNPRED predicts the residue states of the

proteins in the database with 69% accuracy and scores 12 percentage points higher

than a simple statistical method. Moreover the server implements a neural network

to predict the relative solvent accessibility of each residue. A protein sequence

can be directly submitted to RCNPRED: residue co-ordination numbers and solvent

accessibility for each chain are returned via e-mail.AVAILABILITY: Freely

available to non-commercial users at http://prion.biocomp.unibo.it/rcnpred.html.

PMID: 11238082 [Indexed for MEDLINE]

3670. Bioinformatics. 2001 Feb;17(2):126-36.

A hierarchical unsupervised growing neural network for clustering gene expression

patterns.

Herrero J(1), Valencia A, Dopazo J.

Author information:

(1)Bioinformatics, CNIO, Ctra. Majadahonda-Pozuelo, Km 2, Majadahonda, 28220

Madrid Protein Design Group CNB-CSIC, 28049 Madrid, Spain.

MOTIVATION: We describe a new approach to the analysis of gene expression data

coming from DNA array experiments, using an unsupervised neural network. DNA

array technologies allow monitoring thousands of genes rapidly and efficiently.

One of the interests of these studies is the search for correlated gene

expression patterns, and this is usually achieved by clustering them. The

Self-Organising Tree Algorithm, (SOTA) (Dopazo,J. and Carazo,J.M. (1997) J. Mol.

Evol., 44, 226-233), is a neural network that grows adopting the topology of a

binary tree. The result of the algorithm is a hierarchical cluster obtained with

the accuracy and robustness of a neural network.

RESULTS: SOTA clustering confers several advantages over classical hierarchical

clustering methods. SOTA is a divisive method: the clustering process is

performed from top to bottom, i.e. the highest hierarchical levels are resolved

before going to the details of the lowest levels. The growing can be stopped at

the desired hierarchical level. Moreover, a criterion to stop the growing of the

tree, based on the approximate distribution of probability obtained by

randomisation of the original data set, is provided. By means of this criterion,

a statistical support for the definition of clusters is proposed. In addition,

obtaining average gene expression patterns is a built-in feature of the

algorithm. Different neurons defining the different hierarchical levels represent

the averages of the gene expression patterns contained in the clusters. Since

SOTA runtimes are approximately linear with the number of items to be classified,

it is especially suitable for dealing with huge amounts of data. The method

proposed is very general and applies to any data providing that they can be coded

as a series of numbers and that a computable measure of similarity between data

items can be used.

AVAILABILITY: A server running the program can be found at:

http://bioinfo.cnio.es/sotarray.

PMID: 11238068 [Indexed for MEDLINE]

3671. Protein Sci. 2001 Feb;10(2):352-61.

LiveBench-1: continuous benchmarking of protein structure prediction servers.

Bujnicki JM(1), Elofsson A, Fischer D, Rychlewski L.

Author information:

(1)Bioinformatics Laboratory, International Institute of Molecular and Cell

Biology, 02-109 Warsaw, Poland.

We present a novel, continuous approach aimed at the large-scale assessment of

the performance of available fold-recognition servers. Six popular servers were

investigated: PDB-Blast, FFAS, T98-lib, GenTHREADER, 3D-PSSM, and INBGU. The

assessment was conducted using as prediction targets a large number of selected

protein structures released from October 1999 to April 2000. A target was

selected if its sequence showed no significant similarity to any of the proteins

previously available in the structural database. Overall, the servers were able

to produce structurally similar models for one-half of the targets, but

significantly accurate sequence-structure alignments were produced for only

one-third of the targets. We further classified the targets into two sets: easy

and hard. We found that all servers were able to find the correct answer for the

vast majority of the easy targets if a structurally similar fold was present in

the server's fold libraries. However, among the hard targets--where standard

methods such as PSI-BLAST fail--the most sensitive fold-recognition servers were

able to produce similar models for only 40% of the cases, half of which had a

significantly accurate sequence-structure alignment. Among the hard targets, the

presence of updated libraries appeared to be less critical for the ranking. An

"ideally combined consensus" prediction, where the results of all servers are

considered, would increase the percentage of correct assignments by 50%. Each

server had a number of cases with a correct assignment, where the assignments of

all the other servers were wrong. This emphasizes the benefits of considering

more than one server in difficult prediction tasks. The LiveBench program

(http://BioInfo.PL/LiveBench) is being continued, and all interested developers

are cordially invited to join.

DOI: 10.1110/ps.40501

PMCID: PMC2373940

PMID: 11266621 [Indexed for MEDLINE]

3672. Bioinformatics. 2001;17 Suppl 1:S234-42.

Improved prediction of the number of residue contacts in proteins by recurrent

neural networks.

Pollastri G(1), Baldi P, Fariselli P, Casadio R.

Author information:

(1)Department of Information and Computer Science, Institute for Genomics and

Bioinformatics, University of California, Irvine, Irvine, CA 92697-3425, USA.

Knowing the number of residue contacts in a protein is crucial for deriving

constraints useful in modeling protein folding, protein structure, and/or scoring

remote homology searches. Here we use an ensemble of bi-directional recurrent

neural network architectures and evolutionary information to improve the

state-of-the-art in contact prediction using a large corpus of curated data. The

ensemble is used to discriminate between two different states of residue

contacts, characterized by a contact number higher or lower than the average

value of the residue distribution. The ensemble achieves performances ranging

from 70.1% to 73.1% depending on the radius adopted to discriminate contacts

(6Ato 12A). These performances represent gains of 15% to 20% over the base line

statistical predictors always assigning an aminoacid to the most numerous state,

3% to 7% better than any previous method. Combination of different radius

predictors further improves the performance. SERVER:

http://promoter.ics.uci.edu/BRNN-PRED/.

PMID: 11473014 [Indexed for MEDLINE]

3673. BMC Bioinformatics. 2001;2:7. Epub 2001 Oct 10.

The distributed annotation system.

Dowell RD(1), Jokerst RM, Day A, Eddy SR, Stein L.

Author information:

(1)Howard Hughes Medical Institute and Department of Genetics, Washington

University, St. Louis, MO 63110 USA. robin@genetics.wustl.edu

BACKGROUND: Currently, most genome annotation is curated by centralized groups

with limited resources. Efforts to share annotations transparently among multiple

groups have not yet been satisfactory.

RESULTS: Here we introduce a concept called the Distributed Annotation System

(DAS). DAS allows sequence annotations to be decentralized among multiple

third-party annotators and integrated on an as-needed basis by client-side

software. The communication between client and servers in DAS is defined by the

DAS XML specification. Annotations are displayed in layers, one per server. Any

client or server adhering to the DAS XML specification can participate in the

system; we describe a simple prototype client and server example.

CONCLUSIONS: The DAS specification is being used experimentally by Ensembl,

WormBase, and the Berkeley Drosophila Genome Project. Continued success will

depend on the readiness of the research community to adopt DAS and provide

annotations. All components are freely available from the project website

http://www.biodas.org/.

PMCID: PMC58584

PMID: 11667947 [Indexed for MEDLINE]

3674. Genome Inform. 2001;12:184-93.

A mini-greedy algorithm for faster structural RNA stem-loop search.

Gorodkin J(1), Lyngso RB, Stormo GD.

Author information:

(1)Bioinformatics Research Center and Department of Genetics and Ecology,

University of Aarhus, Building 540, Ny Munkegade, DK-8000 Aarhus, Denmark.

gorodkin@bioinf.au.dk

When a set of coregulated genes share a common structural RNA motif, e.g. a

hairpin, most motif search approaches fail to locate the covarying but

structurally conserved motif. There do exist methods that can locate structural

RNA motifs, like FOLDALIGN, but the main problem with these methods is that they

are computationally expensive. In FOLDALIGN, a major contribution to this is the

use of a greedy algorithm to construct the multiple alignment. To ensure good

quality many redundant computations must be made. However, by applying the greedy

algorithm on a carefully selected subset of sequences, near full greedy quality

can be obtained. The basic idea is to estimate the order in which the sequences

entered a good greedy alignment. If such a ranking, found from all pairwise

alignments, is in good agreement with the order of appearance in the multiple

alignment, the core structural motif can be found by performing the greedy

algorithm on just the top sequences in the ranking. The ranking used in this

mini-greedy algorithm is found by using two complementing approaches: 1) When

interpreting the FOLDALIGN score as an inner product (kernel), the sequences can

be ranked according to their distance to their center of mass; 2) We construct an

algorithm that attempts to find the K closest sequences in the vector space

associated with the inner product, and the remaining sequences can be ranked by

their minimum distance to any of the sequences, or to the center of mass in this

set. The two approaches arecompared and merged, and the results discussed. We

also show that structural alignments of near full greedy quality can found in

significantly reduced time, using these methods. The algorithm is being included

in the SLASH (Stem-Loop Align SearcH) server available at

http://www.bioinf.au.dk/slash.

PMID: 11791237 [Indexed for MEDLINE]

3675. J Chem Inf Comput Sci. 2001 Jan-Feb;41(1):100-7.

CheD: chemical database compilation tool, Internet server, and client for SQL

servers.

Trepalin SV(1), Yarkov AV.

Author information:

(1)Institute of Physiologically Active Compounds, Moscow Region, Russia.

trep@ism.ac.ru

An efficient program, which runs on a personal computer, for the storage,

retrieval, and processing of chemical information, is presented, The program can

work both as a stand-alone application or in conjunction with a specifically

written Web server application or with some standard SQL servers, e.g., Oracle,

Interbase, and MS SQL. New types of data fields are introduced, e.g., arrays for

spectral information storage, HTML and database links, and user-defined

functions. CheD has an open architecture; thus, custom data types, controls, and

services may be added. A WWW server application for chemical data retrieval

features an easy and user-friendly installation on Windows NT or 95 platforms.

PMID: 11206361

3676. Nucleic Acids Res. 2001 Jan 1;29(1):173-4.

The RDP-II (Ribosomal Database Project).

Maidak BL(1), Cole JR, Lilburn TG, Parker CT Jr, Saxman PR, Farris RJ, Garrity

GM, Olsen GJ, Schmidt TM, Tiedje JM.

Author information:

(1)Center for Microbial Ecology, 540 Plant and Soil Sciences Building, Michigan

State University, East Lansing, MI 48824-1325, USA.

The Ribosomal Database Project (RDP-II), previously described by Maidak et al.

[Nucleic Acids Res. (2000), 28, 173-174], continued during the past year to add

new rRNA sequences to the aligned data and to improve the analysis commands.

Release 8.0 (June 1, 2000) consisted of 16 277 aligned prokaryotic small subunit

(SSU) rRNA sequences while the number of eukaryotic and mitochondrial SSU rRNA

sequences in aligned form remained at 2055 and 1503, respectively. The number of

prokaryotic SSU rRNA sequences more than doubled from the previous release 14

months earlier, and approximately 75% are longer than 899 bp. An RDP-II mirror

site in Japan is now available (http://wdcm.nig.ac.jp/RDP/html/index.h tml).

RDP-II provides aligned and annotated rRNA sequences, derived phylogenetic trees

and taxonomic hierarchies, and analysis services through its WWW server

(http://rdp.cme.msu.edu/). Analysis services include rRNA probe checking,

approximate phylogenetic placement of user sequences, screening user sequences

for possible chimeric rRNA sequences, automated alignment, production of

similarity matrices and services to plan and analyze terminal restriction

fragment polymorphism experiments. The RDP-II email address for questions and

comments has been changed from curator@cme.msu.edu to rdpstaff@msu.edu.

PMCID: PMC29785

PMID: 11125082 [Indexed for MEDLINE]

3677. Nucleic Acids Res. 2001 Jan 1;29(1):255-9.

SpliceDB: database of canonical and non-canonical mammalian splice sites.

Burset M(1), Seledtsov IA, Solovyev VV.

Author information:

(1)The Sanger Centre, Hinxton, Cambridge CB10 1SA, UK and Softberry Inc., 108

Corporate Park Drive, Suite 120, White Plains, NY 10604, USA.

A database (SpliceDB) of known mammalian splice site sequences has been

developed. We extracted 43 337 splice pairs from mammalian divisions of the

gene-centered Infogene database, including sites from incomplete or alternatively

spliced genes. Known EST sequences supported 22 815 of them. After discarding

sequences with putative errors and ambiguous location of splice junctions the

verified dataset includes 22 489 entries. Of these, 98.71% contain canonical

GT-AG junctions (22 199 entries) and 0.56% have non-canonical GC-AG splice site

pairs. The remainder (0.73%) occurs in a lot of small groups (with a maximum size

of 0.05%). We especially studied non-canonical splice sites, which comprise 3.73%

of GenBank annotated splice pairs. EST alignments allowed us to verify only the

exonic part of splice sites. To check the conservative dinucleotides we compared

sequences of human non-canonical splice sites with sequences from the high

throughput genome sequencing project (HTG). Out of 171 human non-canonical and

EST-supported splice pairs, 156 (91.23%) had a clear match in the human HTG. They

can be classified after sequence analysis as: 79 GC-AG pairs (of which one was an

error that corrected to GC-AG), 61 errors corrected to GT-AG canonical pairs, six

AT-AC pairs (of which two were errors corrected to AT-AC), one case was produced

from a non-existent intron, seven cases were found in HTG that were deposited to

GenBank and finally there were only two other cases left of supported

non-canonical splice pairs. The information about verified splice site sequences

for canonical and non-canonical sites is presented in SpliceDB with the

supporting evidence. We also built weight matrices for the major splice groups,

which can be incorporated into gene prediction programs. SpliceDB is available at

the computational genomic Web server of the Sanger Centre:

http://genomic.sanger.ac. uk/spldb/SpliceDB.html and at http://www.softberry.

com/spldb/SpliceDB.html.

PMCID: PMC29840

PMID: 11125105 [Indexed for MEDLINE]

3678. Nucleic Acids Res. 2001 Jan 1;29(1):219-20.

PDB-REPRDB: a database of representative protein chains from the Protein Data

Bank (PDB).

Noguchi T(1), Matsuda H, Akiyama Y.

Author information:

(1)Computational Biology Research Laboratory, Electrotechnical Laboratory, 1-1-4

Umezono, Tsukuba,Ibaraki 305-5868, Japan. tnoguchi@etl.go.jp

PDB-REPRDB is a database of representative protein chains from the Protein Data

Bank (PDB). The previous version of PDB-REPRDB provided 48 representative sets,

whose similarity criteria were predetermined, on the WWW. The current version is

designed so that the user may obtain a quick selection of representative chains

from PDB. The selection of representative chains can be dynamically configured

according to the user's requirement. The WWW interface provides a large degree of

freedom in setting parameters, such as cut-off scores of sequence and structural

similarity. One can obtain a representative list and classification data of

protein chains from the system. The current database includes 20 457 protein

chains from PDB entries (August 6, 2000). The system for PDB-REPRDB is available

at the Parallel Protein Information Analysis system (PAPIA) WWW server

(http://www.rwcp.or.jp/papia/).

PMCID: PMC29793

PMID: 11125096 [Indexed for MEDLINE]

3679. Nucleic Acids Res. 2001 Jan 1;29(1):175-7.

The European Large Subunit Ribosomal RNA Database.

Wuyts J(1), De Rijk P, Van de Peer Y, Winkelmans T, De Wachter R.

Author information:

(1)Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1,

B-2610 Antwerpen, Belgium.

The European Large Subunit Ribosomal RNA Database compiles all complete or nearly

complete large subunit ribosomal RNA sequences available from public sequence

databases. These are provided in aligned format and the secondary structure, as

derived by comparative sequence analysis, is included. Additional information

about the sequences such as literature references and taxonomic information is

also included. The database is available from our WWW server at

http://rrna.uia.ac.be/lsu/.

PMCID: PMC29789

PMID: 11125083 [Indexed for MEDLINE]

3680. Nucleic Acids Res. 2001 Jan 1;29(1):106-10.

The ARKdb: genome databases for farmed and other animals.

Hu J(1), Mungall C, Law A, Papworth R, Nelson JP, Brown A, Simpson I, Leckie S,

Burt DW, Hillyard AL, Archibald AL.

Author information:

(1)Roslin Institute, Roslin, Midlothian EH25 9PS, Scotland, UK.

The ARKdb genome databases provide comprehensive public repositories for genome

mapping data from farmed species and other animals (http://www.thearkdb.org)

providing a resource similar in function to that offered by GDB or MGD for human

or mouse genome mapping data, respectively. Because we have attempted to build a

generic mapping database, the system has wide utility, particularly for those

species for which development of a specific resource would be prohibitive. The

ARKdb genome database model has been implemented for 10 species to date. These

are pig, chicken, sheep, cattle, horse, deer, tilapia, cat, turkey and salmon.

Access to the ARKdb databases is effected via the World Wide Web using the ARKdb

browser and Anubis map viewer. The information stored includes details of loci,

maps, experimental methods and the source references. Links to other information

sources such as PubMed and EMBL/GenBank are provided. Responsibility for data

entry and curation is shared amongst scientists active in genome research in the

species of interest. Mirror sites in the United States are maintained in addition

to the central genome server at Roslin.

PMCID: PMC29807

PMID: 11125062 [Indexed for MEDLINE]

3681. Nucleic Acids Res. 2001 Jan 1;29(1):49-51.

The MetaFam Server: a comprehensive protein family resource.

Silverstein KA(1), Shoop E, Johnson JE, Kilian A, Freeman JL, Kunau TM, Awad IA,

Mayer M, Retzel EF.

Author information:

(1)Computational Biology Centers, Academic Health Center, University of

Minnesota, Mayo Mail Code 43, 420 Delaware Street, SE Minneapolis, MN 55455-0312,

USA.

MetaFam is a comprehensive relational database of protein family information.

This web-accessible resource integrates data from several primary sequence and

secondary protein family databases. By pooling together the information from

these disparate sources, MetaFam is able to provide the most complete protein

family sets available. Users are able to explore the interrelationships among

these primary and secondary databases using a powerful graphical visualization

tool, MetaFamView. Additionally, users can identify corresponding sequence

entries among the sequence databases, obtain a quick summary of corresponding

families (and their sequence members) among the family databases, and even

attempt to classify their own unassigned sequences. Hypertext links to the

appropriate source databases are provided at every level of navigation. Global

family database statistics and information are also provided. Public access to

the data is available at http://metafam.ahc.umn.edu/.

PMCID: PMC29768

PMID: 11125046 [Indexed for MEDLINE]

3682. Proteomics. 2001 Jan;1(1):136-63.

The mouse SWISS-2D PAGE database: a tool for proteomics study of diabetes and

obesity.

Sanchez JC(1), Chiappe D, Converset V, Hoogland C, Binz PA, Paesano S, Appel RD,

Wang S, Sennitt M, Nolan A, Cawthorne MA, Hochstrasser DF.

Author information:

(1)Clinical Chemistry Laboratory, University Hospital, Geneva, Switzerland.

sanchez@dim.hcuge.ch

A number of two-dimensional electrophoresis (2-DE) reference maps from mouse

samples have been established and could be accessed through the internet. An

up-to-date list can be found in WORLD-2D PAGE (http://www.expasy.ch/ch2d/2d-

index.html), an index of 2-DE databases and services. None of them were

established from mouse white and brown adipose tissues, pancreatic islets, liver

nuclei and skeletal muscle. This publication describes the mouse SWISS-2D PAGE

database. Proteins present in samples of mouse (C57BI/6J) liver, liver nuclei,

muscle, white and brown adipose tissue and pancreatic islets are assembled and

described in an accessible uniform format. SWISS-2D PAGE can be accessed through

the World Wide Web (WWW) network on the ExPASy molecular biology server

(http://www.expasy.ch/ ch2d/).

DOI: 10.1002/1615-9861(200101)1:1<136::AID-PROT136>3.0.CO;2-1

PMID: 11680894 [Indexed for MEDLINE]

3683. Stud Health Technol Inform. 2001;84(Pt 1):715-8.

Migration of the Japanese healthcare enterprise from a financial to integrated

management: strategy and architecture.

Akiyama M(1).

Author information:

(1)Department of Medical Informatics, International Medical Center of Japan,

Tokyo, 162-8655, Japan. makiyama@imcj.hosp.go.jp

The Hospital Information System (HIS) has been positioned as the hub of the

healthcare information management architecture. In Japan, the billing system

assigns an "insurance disease names" to performed exams based on the diagnosis

type. Departmental systems provide localized, departmental services, such as

order receipt and diagnostic reporting, but do not provide patient demographic

information. The system above has many problems. The departmental system's

terminals and the HIS's terminals are not integrated. Duplicate data entry

introduces errors and increases workloads. Order and exam data managed by the HIS

can be sent to the billing system, but departmental data cannot usually be

entered. Additionally, billing systems usually keep departmental data for only a

short time before it is deleted. The billing system provides payment based on

what is entered. The billing system is oriented towards diagnoses. Most

importantly, the system is geared towards generating billing reports rather than

at providing high-quality patient care. The role of the application server is

that of a mediator between system components. Data and events generated by system

components are sent to the application server that routes them to appropriate

destinations. It also records all system events, including state changes to

clinical data, access of clinical data and so on. Finally, the Resource

Management System identifies all system resources available to the enterprise.

The departmental systems are responsible for managing data and clinical processes

at a departmental level. The client interacts with the system via the application

server, which provides a general set of system-level functions. The system is

implemented using current technologies CORBA and HTTP. System data is collected

by the application server and assembled into XML documents for delivery to

clients. Clients can access these URLs using standard HTTP clients, since each

department provides an HTTP compliant web-server. We have implemented an

integrated system communicating via CORBA middleware, consisting of an

application server, endoscopy departmental server, pathology departmental server

and wrappered legacy HIS. We have found this new approach solves the problems

outlined earlier. It provides the services needed to ensure that data is never

lost and is always available, that events that occur in the hospital are always

captured, and that resources are managed and tracked effectively. Finally, it

reduces costs, raises efficiency, increases the quality of patient care, and

ultimately saves lives. Now, we are going to integrate all remaining hospital

departments, and ultimately, all hospital functions.

PMID: 11604829 [Indexed for MEDLINE]

3684. Verh Dtsch Ges Pathol. 2001;85:72-9.

[Teaching of histopathology on the Internet].

[Article in German]

Herbst H(1), Hübner JH.

Author information:

(1)Gerhard-Domagk-Institut für Pathologie, Westfälische Wilhelms-Universität

Münster.

The authors report their experience with the German language histopathology

teaching web site www.pathologie-online.de, and discuss its design, web server

statistics and future developments. The current status of German language

internet resources designed for teaching of pathology at medical schools is

reviewed and compared to corresponding English language resources with public

access.

PMID: 11894417 [Indexed for MEDLINE]

3685. Z Kardiol. 2001 Jan;90(1):12-20.

[Standardized findings in echocardiography using WWW: EchoBefundSystem].

[Article in German]

Schweikart O(1), Metzger F.

Author information:

(1)DFKI GmbH Erwin-Schrödinger-Str. Geb. 57 67663 Kaiserslautern.

schweik@dfki-uni.kl.de bzw Oliver.Schweikart@web.de

As a non-invasive imaging system, ultrasound echocardiography has a very high

impact on modern diagnosis and is widely used in clinical routine but without any

structured and standardized documentation of the results. Thus, quality

management (QM), statistics and comparison of the results are difficult.

Therefore, a working group of the German Cardiac Society issued a consensus

proposal. For evaluation and wide public distribution, we have developed the

first internet-based application covering the full proposal: EchoBefundSystem.

The EchoBefundSystem is a web based client-server application for standardized

documentation of echocardiography results right at the workplace. The software

leads the examiner by means of a user interface design and stored medical

knowledge. The level of detail is scaled automatically to the ongoing

examination. Every day clinical routine is performed on only two pages, one for

general patient data and a second one covering the complete minimal data set

called "minimum finding" in the standard. As the examiner discovers more and more

special findings or might even enter a complete medical study, the interface

offers more and more fields and checkboxes. One data set can contain up to 600

values and findings. The structured user interface reflects the organ structure

as well as examination methods familiar to the examiner. Automatically calculated

fields speed up the examination. Judgements, diagnoses, values and ranges are

interrelated. If there is a difference between the entered data and the medical

knowledge base, a warning will be issued but the doctor's decision is

authoritative. Some values may be gathered by different methods and even

different units are converted automatically. The final doctor's letter is

generated automatically in clear text but still editable and can be given out to

the patient right after the examination without any further delay. Beside the

minimal data set, all abnormal findings will appear and findings will be

summarized wherever possible. The report is intended for the referring general

practitioner, your own documentation, expert witness as well as clinical studies.

Interested examiners may test the full version online at

http://echo.ma.uni-heidelberg.de

PMID: 11220082 [Indexed for MEDLINE]

3686. J Med Syst. 2000 Dec;24(6):333-8.

Accessing endoscopic images for remote conference and diagnosis using WWW server

with a secure socket layer.

Hashiba M(1), Matsuto T, Arai F, Yamakawa T, Akazawa K.

Author information:

(1)Department of Medical Informatics, Niigata University Medical Hospital,

Niigata University, Niigata, Japan.

Out of 137 hospitals in Niigata prefecture, 13 are connected to the Internet. The

number of private clinics using the Internet has also increased. It is thought

that an endoscopic information exchange over the Internet connection is useful

for the cooperation between hospitals, clinics and organizations in related

fields. A conventional World Wide Web (WWW) server of endoscopic images has been

built around a compressed JPEG (Joint Panel Expert Group) format, for remote

conferences and diagnoses. It has been made secure using a proxy server, a

firewall server and a Secure Socket Layer (SSL). It is easy to access the server

and view these endoscopic images using the usual homepage browsing method.

Ordinarily, gastroenterologists would diagnose lesions with endoscopy in most

cases. It is suggested that this conventional WWW server will be effective for

remote conferences in some cases, even consultation will be possible.

PMID: 11143591 [Indexed for MEDLINE]

3687. Bioinformatics. 2000 Nov;16(11):1044-5.

Model.it: building three dimensional DNA models from sequence data.

Vlahovicek K(1), Pongor S.

Author information:

(1)Protein Structure and Function Group, International Centre for Genetic

Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy.

kristian@icgeb.trieste.it

A WWW server is described for creating 3D models of canonical or bent DNA

starting from sequence data. Predicted DNA trajectory is first computed based on

a choice of di- and tri-nucleotide models (M.G. Munteanu et al., Trends Biochem.

Sci. 23, 341-347, 1998); an atomic model is then constructed and optionally

energy-minimized with constrained molecular dynamics. The data are presented as a

standard PDB file, directly viewable on the user's PC using any molecule

manipulation program.AVAILABILITY: The model.it server is freely available at

http://www.icgeb.trieste.it/dna/

CONTACT: kristian@icgeb.trieste.it; pongor@icgeb.trieste.it

SUPPLEMENTARY INFORMATION: a series of help files is available at the above

address.

PMID: 11269231 [Indexed for MEDLINE]

3688. Bioinformatics. 2000 Nov;16(11):1046-7.

VISTA : visualizing global DNA sequence alignments of arbitrary length.

Mayor C(1), Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS,

Dubchak I.

Author information:

(1)National Energy Research Scientific Computing Center Genome Sciences

Department, Berkeley, CA 94720, USA. vista@lbl.gov

SUMMARY: VISTA is a program for visualizing global DNA sequence alignments of

arbitrary length. It has a clean output, allowing for easy identification of

similarity, and is easily configurable, enabling the visualization of alignments

of various lengths at different levels of resolution. It is currently available

on the web, thus allowing for easy access by all researchers.

AVAILABILITY: VISTA server is available on the web at

http://www-gsd.lbl.gov/vista. The source code is available upon request.

CONTACT: vista@lbl.gov

PMID: 11159318 [Indexed for MEDLINE]

3689. Bioinformatics. 2000 Oct;16(10):948-9.

A space-efficient algorithm for aligning large genomic sequences.

Morgenstern B(1).

Author information:

(1)MIPS-Max-Planck-Institut für Biochemic, Am Klopferspitz 18a, 82152

Martinsried, Germany. morgenstern@mips.biochem.mpg.de

SUMMARY: In the segment-by-segment approach to sequence alignment, pairwise and

multiple alignments are generated by comparing gap-free segments of the sequences

under study. This method is particularly efficient in detecting local homologies,

and it has been used to identify functional regions in large genomic sequences.

Herein, an algorithm is outlined that calculates optimal pairwise

segment-by-segment alignments in essentially linear space. AVAILABILTIY: The

program is available at the Bielefeld Bioinformatics Server (BiBiServ) at

http://bibiserv.techfak. uni-bielefeld.de/dialign/

PMID: 11120687 [Indexed for MEDLINE]

3690. Bioinformatics. 2000 Oct;16(10):915-22.

CAST: an iterative algorithm for the complexity analysis of sequence tracts.

Complexity analysis of sequence tracts.

Promponas VJ(1), Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C,

Ouzounis CA.

Author information:

(1)Department of Cell Biology and Biophysics, Faculty of Biology, University of

Athens, Athens GR-15701, Greece.

MOTIVATION: Sensitive detection and masking of low-complexity regions in protein

sequences. Filtered sequences can be used in sequence comparison without the risk

of matching compositionally biased regions. The main advantage of the method over

similar approaches is the selective masking of single residue types without

affecting other, possibly important, regions.

RESULTS: A novel algorithm for low-complexity region detection and selective

masking. The algorithm is based on multiple-pass Smith-Waterman comparison of the

query sequence against twenty homopolymers with infinite gap penalties. The

output of the algorithm is both the masked query sequence for further analysis,

e.g. database searches, as well as the regions of low complexity. The detection

of low-complexity regions is highly specific for single residue types. It is

shown that this approach is sufficient for masking database query sequences

without generating false positives. The algorithm is benchmarked against widely

available algorithms using the 210 genes of Plasmodium falciparum chromosome 2, a

dataset known to contain a large number of low-complexity regions.

AVAILABILITY: CAST (version 1.0) executable binaries are available to academic

users free of charge under license. Web site entry point, server and additional

material: http://www.ebi.ac.uk/research/cgg/services/cast/

PMID: 11120681 [Indexed for MEDLINE]

3691. Bioinformatics. 2000 Oct;16(10):899-905.

USAGE: a web-based approach towards the analysis of SAGE data. Serial Analysis of

Gene Expression.

van Kampen AH(1), van Schaik BD, Pauws E, Michiels EM, Ruijter JM, Caron HN,

Versteeg R, Heisterkamp SH, Leunissen JA, Baas F, van der Mee M.

Author information:

(1)Bioinformatics Laboratory, Academic Medical Center, Meibergdreef 9, 1000 AZ

Amsterdam, The Netherlands. a.h.vancampen@amc.uva.nl

MOTIVATION: SAGE enables the determination of genome-wide mRNA expression

profiles. A comprehensive analysis of SAGE data requires software, which

integrates (statistical) data analysis methods with a database system.

Furthermore, to facilitate data sharing between users, the application should

reside on a central server and be accessed via the internet. Since such an

application was not available we developed the USAGE package.

RESULTS: USAGE is a web-based application that comprises an integrated set of

tools, which offers many functions for analysing and comparing SAGE data.

Additionally, USAGE includes a statistical method for the planning of new SAGE

experiments. USAGE is available in a multi-user environment giving users the

option of sharing data. USAGE is interfaced to a relational database to store

data and analysis results. The USAGE query editor allows the composition of

queries for searching this database. Several database functions have been

included which enable the selection and combination of data. USAGE provides the

biologist increased functionality and flexibility for analysing SAGE data.

AVAILABILITY: USAGE is freely accessible for academic institutions at

http://www.cmbi.kun.nl/usage/. The source code of USAGE is freely available for

academic institutions on request from the first author.

PMID: 11120679 [Indexed for MEDLINE]

3692. Electrophoresis. 2000 Oct;21(16):3483-7.

The establishment of a human liver nuclei two-dimensional electrophoresis

reference map.

Jung E(1), Hoogland C, Chiappe D, Sanchez JC, Hochstrasser DF.

Author information:

(1)Central Clinical Chemistry Laboratory, Geneva University Hospital,

Switzerland. eva.jung@dim.hcuge.ch

This short communication describes the establishment of a two-dimensional

electrophoresis (2-DE) reference map of nuclear proteins isolated from human

liver. The human liver nuclei 2-DE reference map contains 1497 spots. In an

initial identification study using peptide mass fingerprinting as a means of

protein identification we were able to identify 26 spots corresponding to 15

different proteins. The human liver nuclei 2-DE reference map is now included in

the SWISS-2DPAGE database, which can be accessed through the ExPASy server

(http://www.expasy.ch/ch2d/).

DOI: 10.1002/1522-2683(20001001)21:16<3483::AID-ELPS3483>3.0.CO;2-X

PMID: 11079567 [Indexed for MEDLINE]

3693. Health Libr Rev. 2000 Sep;17(3):164-70.

Local information for primary care: the St Albans WAX Project (STAPCIS).

Rousseau N.

The St. Albans Primary Care Information Service (STAPCIS) is a database of

locally focused information designed to be of use for GPs and the Primary Care

team as a whole in a particular area. It is currently being used by practices in

the St. Albans, Dacorum and Watford & Three Rivers Primary Care Groups. STAPCIS

developed from a project initiated in late 1997 to pilot the use of the WAX

software to improve access to directory-type and full-text material needed for

easy access by primary care staff and traditionally difficult to manage. The

database is updated centrally by the STAPCIS Librarian who distributes the new

editions to the practices. The WAX software itself was developed by the Cambridge

Centre for Clinical Informatics. STAPCIS contains locally and nationally produced

relevant information from a variety of sources, including full-text clinical

guidelines and directories of Trusts. Decisions on new content are made by a

Steering Group and the GPs and Primary Care teams are particularly encouraged to

make suggestions regarding new content. The service is financially supported by

the three primary care groups. The STAPCIS Librarian is project-managed by the

Library & Information Development Unit (North Thames

<http:¿www.nthameshealth.tpmde.ac.uk/rliu/ ntrliu.htm>), which also provides

additional equipment and expertise. A new software company, WaX Info Ltd

(<http:¿www.waxinfo.com/>), will soon be releasing a new version of the software,

WaX. WaX consists of a WaX client and WaX ActiveLibrary server software. With the

new version, the 'books' of information will still be stored on the user's PC,

retaining the speed of access, but the server will enable users to 'borrow' books

from remote locations and will provide automatic updates of all books. The new

software will improve the management and distribution processes of STAPCIS once

all of the practices have Internet connectivity.

PMID: 11186809 [Indexed for MEDLINE]

3694. IEEE Trans Inf Technol Biomed. 2000 Sep;4(3):212-5.

A WEB-based telePACS using an asymmetric satellite system.

Hwang SC(1), Lee MH.

Author information:

(1)Department of Broadcasting and Communication, Induk Institute of Technology,

Wolgae Nowon, Seoul, Korea.

We have developed a WWW-based TelePACS that can access every permitted PACS

server via the Internet. Java programming techniques were used to implement the

system, which can access and retrieve medical information and images through Web

browsers only such as Netscape without specific tools. We also have developed a

consolidator that performs as a manager to connect a conventional PACS server to

a Web-based TelePACS server. We have developed the Asymmetric Satellite Data

Communication System (ASDCS) as a fast communication system. The ASDCS uses a

receive-only satellite link for data delivery and a terrestrial network for

control communication. In conclusion, we were able to develop a cost-effective

and fast PACS using Web technology. Web technology expanded the scope of use for

a dedicated PACS from intrahospital to public use.

PMID: 11026591 [Indexed for MEDLINE]

3695. Proteins. 2000 Aug 15;40(3):502-11.

Application of multiple sequence alignment profiles to improve protein secondary

structure prediction.

Cuff JA(1), Barton GJ.

Author information:

(1)Laboratory of Molecular Biophysics, Oxford, United Kingdom.

The effect of training a neural network secondary structure prediction algorithm

with different types of multiple sequence alignment profiles derived from the

same sequences, is shown to provide a range of accuracy from 70.5% to 76.4%. The

best accuracy of 76.4% (standard deviation 8.4%), is 3.1% (Q(3)) and 4.4% (SOV2)

better than the PHD algorithm run on the same set of 406 sequence non-redundant

proteins that were not used to train either method. Residues predicted by the new

method with a confidence value of 5 or greater, have an average Q(3) accuracy of

84%, and cover 68% of the residues. Relative solvent accessibility based on a two

state model, for 25, 5, and 0% accessibility are predicted at 76.2, 79.8, and 86.

6% accuracy respectively. The source of the improvements obtained from training

with different representations of the same alignment data are described in

detail. The new Jnet prediction method resulting from this study is available in

the Jpred secondary structure prediction server, and as a stand-alone computer

program from: http://barton.ebi.ac.uk/. Proteins 2000;40:502-511.

Copyright 2000 Wiley-Liss, Inc.

PMID: 10861942 [Indexed for MEDLINE]

3696. Bioinformatics. 2000 Aug;16(8):747-8.

PASS: prediction of activity spectra for biologically active substances.

Lagunin A(1), Stepanchikova A, Filimonov D, Poroikov V.

Author information:

(1)Laboratory of Structure-Function Based Drug Design, Institute of Biomedical

Chemistry RAMS, Moscow, Pogodinskaya str., 10, 119832, Russia. pass@ibmh.msk.su

The concept of the biological activity spectrum was introduced to describe the

properties of biologically active substances. The PASS (prediction of activity

spectra for substances) software product, which predicts more than 300

pharmacological effects and biochemical mechanisms on the basis of the structural

formula of a substance, may be efficiently used to find new targets (mechanisms)

for some ligands and, conversely, to reveal new ligands for some biological

targets. We have developed a WWW interface for the PASS software. A WWW server

for the on-line prediction of the biological activity spectra of substances has

been constructed.

PMID: 11099264 [Indexed for MEDLINE]

3697. Methods Inf Med. 2000 Aug;39(3):267-77.

Recommendations of the International Medical Informatics Association (IMIA) on

education in health and medical informatics.

[No authors listed]

The International Medical Informatics Association (IMIA) agreed on international

recommendations in health informatics/medical informatics education. These should

help to establish courses, course tracks or even complete programs in this field,

to further develop existing educational activities in the various nations and to

support international initiatives concerning education in health and medical

informatics (HMI), particularly international activities in educating HMI

specialists and the sharing of courseware. The IMIA recommendations centre on

educational needs for healthcare professionals to acquire knowledge and skills in

information processing and information and communication technology. The

educational needs are described as a three-dimensional framework. The dimensions

are: 1) professionals in healthcare (physicians, nurses, HMI professionals, ...),

2) type of specialisation in health and medical informatics (IT users, HMI

specialists) and 3) stage of career progression (bachelor, master, ...). Learning

outcomes are defined in terms of knowledge and practical skills for healthcare

professionals in their role (a) as IT user and (b) as HMI specialist.

Recommendations are given for courses/course tracks in HMI as part of educational

programs in medicine, nursing, healthcare management, dentistry, pharmacy, public

health, health record administration, and informatics/computer science as well as

for dedicated programs in HMI (with bachelor, master or doctor degree). To

support education in HMI, IMIA offers to award a certificate for high quality HMI

education and supports information exchange on programs and courses in HMI

through a WWW server of its Working Group on Health and Medical Informatics

Education (http:www.imia.org/wg1).

PMID: 10992757 [Indexed for MEDLINE]

3698. J Mol Biol. 2000 Jul 21;300(4):1005-16.

Predicting subcellular localization of proteins based on their N-terminal amino

acid sequence.

Emanuelsson O(1), Nielsen H, Brunak S, von Heijne G.

Author information:

(1)Stockholm Bioinformatics Center, Department of Biochemistry, Stockholm

University, Stockholm, S-106 91, Sweden.

A neural network-based tool, TargetP, for large-scale subcellular location

prediction of newly identified proteins has been developed. Using N-terminal

sequence information only, it discriminates between proteins destined for the

mitochondrion, the chloroplast, the secretory pathway, and "other" localizations

with a success rate of 85% (plant) or 90% (non-plant) on redundancy-reduced test

sets. From a TargetP analysis of the recently sequenced Arabidopsis thaliana

chromosomes 2 and 4 and the Ensembl Homo sapiens protein set, we estimate that

10% of all plant proteins are mitochondrial and 14% chloroplastic, and that the

abundance of secretory proteins, in both Arabidopsis and Homo, is around 10%.

TargetP also predicts cleavage sites with levels of correctly predicted sites

ranging from approximately 40% to 50% (chloroplastic and mitochondrial

presequences) to above 70% (secretory signal peptides). TargetP is available as a

web-server at http://www.cbs.dtu.dk/services/TargetP/.

Copyright 2000 Academic Press.

DOI: 10.1006/jmbi.2000.3903

PMID: 10891285 [Indexed for MEDLINE]

3699. Electrophoresis. 2000 Jul;21(12):2566-75.

Protein analysis by mass spectrometry and sequence database searching: a

proteomic approach to identify human lymphoblastoid cell line proteins.

Joubert-Caron R(1), Le Caër JP, Montandon F, Poirier F, Pontet M, Imam N,

Feuillard J, Bladier D, Rossier J, Caron M.

Author information:

(1)Biochimie Cellulaire des Hémopathies Lymphoïdes, Université Paris 13, UFR

SMBH, Leonard de Vinici, Bibogny, France. caron@smbh.univ-paris13.fr

Lymphoblastoid cell lines correspond to in vitro EBV-immortalized lymphocyte

B-cells. These cells display a suitable model for experiments dealing with

changes in protein expression occurring upon B-cell differentiation, after drug

treatment, or after inhibition of some transcription factors. For all these

reasons we have undertaken an effort aimed at developing a hematopoietic cell

line protein two-dimensional electrophoresis (2-DE) database, containing

B-lymphoblastoid 2-DE maps. In this work, matrix-assisted laser

desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) peptide

mass fingerprinting analysis was adopted for protein identification. The peptide

mass fingerprinting identification and the sequence coverage obtained on

colloidal Coomassie blue (CBB) stained gel was close to that obtained using

zinc-imidazole staining. Everything considered, CBB being more comfortable for

subsequent spot manipulations, CBB staining was chosen for identification of a

larger number of polypeptides. The results suggest that reticulation of the gel

can interfere preventing the uptake of the enzyme during the in-gel digestion

step. Consequently, low molecular mass proteins appear more difficult to identify

by mass fingerprinting. Finally, the information provided in this study allows

the construction of a new annoted reference map of human lymphoblastoid cell

proteins. Among the identified proteins 60% were not yet positioned on 2-DE maps

in three of the most important well-documented databases. The annoted map will be

accessible via Internet on the LBPP server at URL:http://

www-smbh.univ-paris13.fr/lbtp/index.htm.

DOI: 10.1002/1522-2683(20000701)21:12<2566::AID-ELPS2566>3.0.CO;2-F

PMID: 10939474 [Indexed for MEDLINE]

3700. J Mol Biol. 2000 Jun 2;299(2):499-520.

Enhanced genome annotation using structural profiles in the program 3D-PSSM.

Kelley LA(1), MacCallum RM, Sternberg MJ.

Author information:

(1)Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's

Inn Fields, London, WC2A 3PX, England.

A method (three-dimensional position-specific scoring matrix, 3D-PSSM) to

recognise remote protein sequence homologues is described. The method combines

the power of multiple sequence profiles with knowledge of protein structure to

provide enhanced recognition and thus functional assignment of newly sequenced

genomes. The method uses structural alignments of homologous proteins of similar

three-dimensional structure in the structural classification of proteins (SCOP)

database to obtain a structural equivalence of residues. These equivalences are

used to extend multiply aligned sequences obtained by standard sequence searches.

The resulting large superfamily-based multiple alignment is converted into a

PSSM. Combined with secondary structure matching and solvation potentials,

3D-PSSM can recognise structural and functional relationships beyond

state-of-the-art sequence methods. In a cross-validated benchmark on 136

homologous relationships unambiguously undetectable by position-specific iterated

basic local alignment search tool (PSI-Blast), 3D-PSSM can confidently assign 18

%. The method was applied to the remaining unassigned regions of the Mycoplasma

genitalium genome and an additional 13 regions were assigned with 95 %

confidence. 3D-PSSM is available to the community as a web server:

http://www.bmm.icnet.uk/servers/3dpssm

Copyright 2000 Academic Press.

DOI: 10.1006/jmbi.2000.3741

PMID: 10860755 [Indexed for MEDLINE]

3701. Bioinformatics. 2000 Jun;16(6):513-9.

Browsing the SLoop database of structurally classified loops connecting elements

of protein secondary structure.

Burke DF(1), Deane CM, Blundell TL.

Author information:

(1)Department of Biochemistry, University of Cambridge, Tennis Court Road,

Cambridge CB1 2GA1, UK.

We describe a web server, which provides easy access to the SLoop database of

loop conformations connecting elements of protein secondary structure. The loops

are classified according to their length, the type of bounding secondary

structures and the conformation of the mainchain. The current release of the

database consists of over 8000 loops of up to 20 residues in length. A loop

prediction method, which selects conformers on the basis of the sequence and the

positions of the elements of secondary structure, is also implemented. These web

pages are freely accessible over the internet at http://www-cryst.bioc.cam.ac.uk/

approximately sloop.

PMID: 10980148 [Indexed for MEDLINE]

3702. J Pathol. 2000 May;191(1):8-14.

Telepathology by the Internet.

Petersen I(1), Wolf G, Roth K, Schlüns K.

Author information:

(1)Institute of Pathology, University Hospital of the Charité, Campus Mitte,

Berlin, Germany. iver.petersen@charite.de

A new concept for telemicroscopy has recently been introduced using the Internet

and conventional web browser, with Java support for microscope remote control as

well as image transfer and discussion (http://amba.charite.de/telemic/). The

system has two major components: the telemicroscopy server, which is a computer

with Internet access connected to the automatic microscope, and the

telemicroscopy client, who remotely operates the microscope. This simplified

telemicroscopy system allows any Internet user to become a consultant for

telepathology without the acquisition of specialized hardware or software. For

the inquirer seeking advice, however, this solution is still very expensive,

since it requires a fully automated microscope. The present study describes a

system that can be used for conventional microscopes. A video camera mounted on a

microscope with a photo tube is connected to the frame grabber of a PC.

Java-based telemicroscopy software transforms the computer into an Internet

server, which automatically distributes new microscope images, after manual

operations, to all connected clients. Any Internet user can access the web page

of the server to become a telemicroscopy client. A Chat function allows for the

online exchange of written text and a Discuss function enables the mouse button

to display an arrow to all connected clients, which highlights distinct

structures of the images. The system was optimized for simplicity, while

presenting all features that are necessary to show and discuss difficult cases

with any expert in the field who has Internet access. It offers new perspectives

for telepathology and it is envisaged that many pathologists and scientists will

use this facility to connect their personal microscopes to the Internet, forming

a network for teleconsultation. To foster this development, the software

described in this paper is being made freely available. Hopefully, this

development will promote communication between pathologists and may thus increase

the quality of diagnosis. Information on inquiry and installation of the software

is available at the website mentioned above. Telemicroscopy sessions using the

Telemic version for conventional microscopes can be scheduled by contacting the

authors by e-mail (iver. petersen@charite.de).

Copyright 2000 John Wiley & Sons, Ltd.

DOI: 10.1002/(SICI)1096-9896(200005)191:1<8::AID-PATH558>3.0.CO;2-9

PMID: 10767712 [Indexed for MEDLINE]

3703. Nucleic Acids Res. 2000 Apr 15;28(8):1665-75.

The morph server: a standardized system for analyzing and visualizing

macromolecular motions in a database framework.

Krebs WG(1), Gerstein M.

Author information:

(1)Department of Molecular Biophysics and Biochemistry, Yale University, PO Box

208114, New Haven, CT 06520, USA.

The number of solved structures of macromolecules that have the same fold and

thus exhibit some degree of conformational variability is rapidly increasing. It

is consequently advantageous to develop a standardized terminology for describing

this variability and automated systems for processing protein structures in

different conformations. We have developed such a system as a 'front-end' server

to our database of macromolecular motions. Our system attempts to describe a

protein motion as a rigid-body rotation of a small 'core' relative to a larger

one, using a set of hinges. The motion is placed in a standardized coordinate

system so that all statistics between any two motions are directly comparable. We

find that while this model can accommodate most protein motions, it cannot

accommodate all; the degree to which a motion can be accommodated provides an aid

in classifying it. Furthermore, we perform an adiabatic mapping (a restrained

interpolation) between every two conformations. This gives some indication of the

extent of the energetic barriers that need to be surmounted in the motion, and as

a by-product results in a 'morph movie'. We make these movies available over the

Web to aid in visualization. Many instances of conformational variability occur

between proteins with somewhat different sequences. We can accommodate these

differences in a rough fashion, generating an 'evolutionary morph'. Users have

already submitted hundreds of examples of protein motions to our server,

producing a comprehensive set of statistics. So far the statistics show that the

median submitted motion has a rotation of approximately 10 degrees and a maximum

Calpha displacement of 17 A. Almost all involve at least one large torsion angle

change of >140 degrees. The server is accessible at http://bioinfo.mbb.yale.

edu/MolMovDB

PMCID: PMC102811

PMID: 10734184 [Indexed for MEDLINE]

3704. Bioinformatics. 2000 Apr;16(4):404-5.

The PSIPRED protein structure prediction server.

McGuffin LJ(1), Bryson K, Jones DT.

Author information:

(1)Protein Bioinformatics Group, Department of Biological Sciences, University of

Warwick, Coventry CV4 7AL, UK.

SUMMARY: The PSIPRED protein structure prediction server allows users to submit a

protein sequence, perform a prediction of their choice and receive the results of

the prediction both textually via e-mail and graphically via the web. The user

may select one of three prediction methods to apply to their sequence: PSIPRED, a

highly accurate secondary structure prediction method; MEMSAT 2, a new version of

a widely used transmembrane topology prediction method; or GenTHREADER, a

sequence profile based fold recognition method.

AVAILABILITY: Freely available to non-commercial users at

http://globin.bio.warwick.ac.uk/psipred/

PMID: 10869041 [Indexed for MEDLINE]

3705. Bioinformatics. 2000 Apr;16(4):301-12.

GABAagent: a system for integrating data on GABA receptors.

Rachedi A(1), Rebhan M, Xue H.

Author information:

(1)Department of Biochemistry, Hong Kong University of Science and Technology,

Clear Water Bay, Kowloon, Hong Kong.

MOTIVATION: Scientific data pertaining to GABA receptors, which are of medical

importance, are widely scattered throughout numerous heterogeneous Internet

resources. This situation has made the integrated acquisition of such data

difficult and substantially time consuming even for researchers who are Internet

aficionados. Thus, there exists a genuine need for the development of Internet

applications, such as GABAagent, which provide efficient and timely access to

concise and integrated information.

RESULTS: We report here the establishment of a novel server (GABAagent) which has

been written in Perl script, and which is freely accessible through the Internet.

GABAagent is designed to assist researchers in retrieving focused and integrated

information related to GABA receptors from various public domain databases.

GABAagent relies on server-side flat-file databases that have been created

through data mining from Internet sources such as the PubMed, DDBJ, SWISS-PROT

and TrEMBL, in addition to the many World Wide Web (Web) sites which are

accessible through Excite (E-Web). These warehouse databases are regularly

updated and contain among other things, information concerning: (i) GABA receptor

publications, (ii) DNA and protein sequences and (iii) the contents of related

E-Web sites along with their addresses. Our system also provides hard links to

the above-mentioned Web sites and E-Web sites; the feature which adds to it the

character of virtual federation type of database. The current version of

GABAagent provides two user-friendly services. The first is a search engine

possessing intelligent query reformulation support (GABAengine), the second an

elaborate email alert service was designed into the system (GABAalert). The

GABAengine allows the user to search server-side databases exclusively for GABA

receptor-related queries. Whereas, GABAalert allows the user, by means of

subscription, to receive immediate and/or monthly updates automatically.

AVAILABILITY: GABAagent is freely accessible at the following Web address

http://www.ust.hk/gaba.

PMID: 10869028 [Indexed for MEDLINE]

3706. Genome Res. 2000 Apr;10(4):577-86.

PipMaker--a web server for aligning two genomic DNA sequences.

Schwartz S(1), Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison

R, Miller W.

Author information:

(1)Department of Computer Science and Engineering, The Pennsylvania State

University, University Park 16802, USA.

PipMaker (http://bio.cse.psu.edu) is a World-Wide Web site for comparing two long

DNA sequences to identify conserved segments and for producing informative,

high-resolution displays of the resulting alignments. One display is a percent

identity plot (pip), which shows both the position in one sequence and the degree

of similarity for each aligning segment between the two sequences in a compact

and easily understandable form. Positions along the horizontal axis can be

labeled with features such as exons of genes and repetitive elements, and colors

can be used to clarify and enhance the display. The web site also provides a plot

of the locations of those segments in both species (similar to a dot plot).

PipMaker is appropriate for comparing genomic sequences from any two related

species, although the types of information that can be inferred (e.g.,

protein-coding regions and cis-regulatory elements) depend on the level of

conservation and the time and divergence rate since the separation of the

species. Gene regulatory elements are often detectable as similar, noncoding

sequences in species that diverged as much as 100-300 million years ago, such as

humans and mice, Caenorhabditis elegans and C. briggsae, or Escherichia coli and

Salmonella spp. PipMaker supports analysis of unfinished or "working draft"

sequences by permitting one of the two sequences to be in unoriented and

unordered contigs.

PMCID: PMC310868

PMID: 10779500 [Indexed for MEDLINE]

3707. Genome Res. 2000 Apr;10(4):516-22.

Ab initio gene finding in Drosophila genomic DNA.

Salamov AA(1), Solovyev VV.

Author information:

(1)The Sanger Centre, Hinxton, Cambridge CB10 1SA, UK.

Comment in

Genome Res. 2000 Apr;10(4):391-3.

Ab initio gene identification in the genomic sequence of Drosophila melanogaster

was obtained using (human gene predictor) and Fgenesh programs that have

organism-specific parameters for human, Drosophila, plants, yeast, and nematode.

We did not use information about cDNA/EST in most predictions to model a real

situation for finding new genes because information about complete cDNA is often

absent or based on very small partial fragments. We investigated the accuracy of

gene prediction on different levels and designed several schemes to predict an

unambiguous set of genes (annotation CGG1), a set of reliable exons (annotation

CGG2), and the most complete set of exons (annotation CGG3). For 49 genes,

protein products of which have clear homologs in protein databases, predictions

were recomputed by Fgenesh+ program. The first annotation serves as the optimal

computational description of new sequence to be presented in a database. Reliable

exons from the second annotation serve as good candidates for selecting the PCR

primers for experimental work for gene structure verification. Our results shows

that we can identify approximately 90% of coding nucleotides with 20% false

positives. At the exon level we accurately predicted 65% of exons and 89%

including overlapping exons with 49% false positives. Optimizing accuracy of

prediction, we designed a gene identification scheme using Fgenesh, which

provided sensitivity (Sn) = 98% and specificity (Sp) = 86% at the base level, Sn

= 81% (97% including overlapping exons) and Sp = 58% at the exon level and Sn =

72% and Sp = 39% at the gene level (estimating sensitivity on std1 set and

specificity on std3 set). In general, these results showed that computational

gene prediction can be a reliable tool for annotating new genomic sequences,

giving accurate information on 90% of coding sequences with 14% false positives.

However, exact gene prediction (especially at the gene level) needs additional

improvement using gene prediction algorithms. The program was also tested for

predicting genes of human Chromosome 22 (the last variant of Fgenesh can analyze

the whole chromosome sequence). This analysis has demonstrated that the 88% of

manually annotated exons in Chromosome 22 were among the ab initio predicted

exons. The suite of gene identification programs is available through the WWW

server of Computational Genomics Group at http://genomic.sanger.ac.uk/gf. html.

PMCID: PMC310882

PMID: 10779491 [Indexed for MEDLINE]

3708. Bioinformatics. 2000 Mar;16(3):286-7.

MPSA: integrated system for multiple protein sequence analysis with client/server

capabilities.

Blanchet C(1), Combet C, Geourjon C, Deléage G.

Author information:

(1)Institut de Biologie et Chimie des Protéines, BCP-CNRS UPR 412, Laboratoire de

conformation des Protéines, 7, Passage du Vercors, 69 367, Lyon Cedex 07, France.

c.blanchet@ibcp.fr

MPSA is a stand-alone software intended to protein sequence analysis with a high

integration level and Web clients/server capabilities. It provides many methods

and tools, which are integrated into an interactive graphical user interface. It

is available for most Unix/Linux and non-Unix systems. MPSA is able to connect to

a Web server (e.g. http://pbil.ibcp.fr/NPSA) in order to perform large-scale

sequence comparison on up-to-date databanks.AVAILABILITY: Free to academic

http://www.ibcp.fr/mpsa/

CONTACT: c.blanchet@ibcp.fr

PMID: 10869021 [Indexed for MEDLINE]

3709. Bioinformatics. 2000 Mar;16(3):251-6.

Predicting the oxidation state of cysteines by multiple sequence alignment.

Fiser A(1), Simon I.

Author information:

(1)Institute of Enzymology, Biological Research Center, Hungarian Academy of

Sciences, Budapest H-1518, PO Box 7, Hungary. fisera@rockefeller.edu

MOTIVATION: Protein sequences found in databanks usually do not report post

translational covalent modifications such as the oxidation state of cystein (Cys)

residues. Accurate prediction of whether a functionally or structurally important

Cys occurs in the oxidized or thiol form would be helpful for molecular biology

experiments and structure prediction.

RESULTS: A new method is presented for predicting the oxidation state of Cys

residues based on multiple sequence alignments and on the observation that Cys

tends to occur in the same oxidation state within the same protein. The

prediction of the redox state of Cys performs above 82%. The oxidation state of

Cys correlates with the cellular location of the given protein within the cell,

but the correlation is not perfect (up to 70%). We also perform a statistical

analysis of the different redox states of Cys found in secondary structures and

buried positions, and of the secondary structures linked by disulfide bonds. The

results suggest that the natural borderline lies between the different oxidation

states of Cys rather than between the half cystines and cysteins.

AVAILABILITY: A web server implementing the prediction method is available at

http://guitar.rockefeller.edu/approximately andras/cyspred.html

CONTACT: fisera@rockefeller.edu

PMID: 10869018 [Indexed for MEDLINE]

3710. Bioinformatics. 2000 Mar;16(3):245-50.

Adaptive encoding neural networks for the recognition of human signal peptide

cleavage sites.

Jagla B(1), Schuchhardt J.

Author information:

(1)Freie Universität Berlin, Institut für Med./Techn. Physik und Lasermedizin,

Krahmerstrasse 6-10, 12045 Berlin, Germany.

MOTIVATION: Data representation and encoding are essential for classification of

protein sequences with artificial neural networks (ANN). Biophysical properties

are appropriate for low dimensional encoding of protein sequence data. However,

in general there is no a priori knowledge of the relevant properties for

extraction of representative features.

RESULTS: An adaptive encoding artificial neural network (ACN) for recognition of

sequence patterns is described. In this approach parameters for sequence encoding

are optimized within the same process as the weight vectors by an evolutionary

algorithm. The method is applied to the prediction of signal peptide cleavage

sites in human secretory proteins and compared with an established predictor for

signal peptides.

CONCLUSION: Knowledge of physico-chemical properties is not necessary for

training an ACN. The advantage is a low dimensional data representation leading

to computational efficiency, easy evaluation of the detected features, and high

prediction accuracy.

AVAILABILITY: A cleavage site prediction server is located at the Humboldt

University http://itb.biologie.hu-berlin.de/ approximately

jo/sig-cleave/ACNpredictor.cgi

CONTACT: jo@itb.hu-berlin.de; berndj@zedat.fu-berlin.de

PMID: 10869017 [Indexed for MEDLINE]

3711. Genome Res. 2000 Mar;10(3):379-85.

HOBACGEN: database system for comparative genomics in bacteria.

Perrière G(1), Duret L, Gouy M.

Author information:

(1)Laboratoire de Biométrie et Biologie Evolutive, Unité Mixte de Recherche

Centre National de la Recherche Scientifique (UMR CNRS) n( degrees ). 5558,

Université Claude Bernard-Lyon 1, 69622 Villeurbanne Cedex, France.

perriere@biomserv.univ-lyon1.fr

We present here HOBACGEN, a database system devoted to comparative genomics in

bacteria. HOBACGEN contains all available protein genes from bacteria, archaea,

and yeast, taken from SWISS-PROT/TrEMBL and classified into families. It also

includes multiple alignments and phylogenetic trees built from these families.

The database is organized under a client/server architecture with a client

written in Java, which may run on any platform. This client integrates a

graphical interface allowing users to select families according to various

criteria and notably to select homologs common to a given set of taxa. This

interface also allows users to visualize multiple alignments and trees associated

to families. In tree displays, protein gene names are colored according to the

taxonomy of the corresponding organisms. Users may access all information

associated to sequences and multiple alignments by clicking on genes. This

graphic tool thus gives a rapid and simple access to all data required to

interpret homology relationships between genes and distinguish orthologs from

paralogs. Instructions for installation of the client or the server are available

at http://pbil.univ-lyon1. fr/databases/hobacgen.html.

PMCID: PMC311423

PMID: 10720578 [Indexed for MEDLINE]

3712. Proteins. 2000 Mar 1;38(4):428-40.

Protein structure alignment using a genetic algorithm.

Szustakowski JD(1), Weng Z.

Author information:

(1)Boston University, Department of Biomedical Engineering, Massachusetts, USA.

We have developed a novel, fully automatic method for aligning the

three-dimensional structures of two proteins. The basic approach is to first

align the proteins' secondary structure elements and then extend the alignment to

include any equivalent residues found in loops or turns. The initial secondary

structure element alignment is determined by a genetic algorithm. After

refinement of the secondary structure element alignment, the protein backbones

are superposed and a search is performed to identify any additional equivalent

residues in a convergent process. Alignments are evaluated using intramolecular

distance matrices. Alignments can be performed with or without sequential

connectivity constraints. We have applied the method to proteins from several

well-studied families: globins, immunoglobulins, serine proteases, dihydrofolate

reductases, and DNA methyltransferases. Agreement with manually curated

alignments is excellent. A web-based server and additional supporting information

are available at http://engpub1.bu.edu/-josephs.

PMID: 10707029 [Indexed for MEDLINE]

3713. Bioinformatics. 2000 Feb;16(2):125-9.

SAWTED: structure assignment with text description--enhanced detection of remote

homologues with automated SWISS-PROT annotation comparisons.

MacCallum RM(1), Kelley LA, Sternberg MJ.

Author information:

(1)Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, London, UK.

R.MacCallum@icrf.icnet.uk

MOTIVATION: Sequence database search methods often identify putative

sub-threshold hits of known function or structure for a given query sequence. It

is widespread practice to filter these hits by hand using knowledge of function

and other factors; to the expert, some hits may appear more sensible than others.

SAWTED (Structure Assignment With Text Description) is an automated solution to

this post-filtering problem which will be applicable to large scale genome

assignments.

RESULTS: A standard document comparison algorithm is applied to text descriptions

extracted from SWISS-PROT annotations. The added value of SAWTED in combination

with PSI-BLAST has been shown with a benchmark of difficult remote homologues

taken from the SCOP structure database.

AVAILABILITY: A WAWTED PSI-BLAST Web server is available to perform sensitive

searches against the protein structure database

(http://www.bmm.icnet.uk/servers/sawted).

CONTACT: R.MacCallum@icrf.icnet.uk

PMID: 10842733 [Indexed for MEDLINE]

3714. Semin Pediatr Surg. 2000 Feb;9(1):11-8.

Internet resources and web pages for pediatric surgeons.

Lugo-Vicente H(1).

Author information:

(1)Department of Pediatric Surgery, University Children's Hospital, and

University of Puerto Rico School of Medicine, Rio Piedras, San Juan 00922, USA.

The Internet, the largest network of connected computers, provides immediate,

dynamic, and downloadable information. By re-architecturing the work place and

becoming familiar with Internet resources, pediatric surgeons have anticipated

the informatics capabilities of this computer-based technology creating a new

vision of work and organization in such areas as patient care, teaching, and

research. This review aims to highlight how Internet navigational technology can

be a useful educational resource in pediatric surgery, examines web pages of

interest, and defines ideas of network communication. Basic Internet resources

are electronic mail, discussion groups, file transfer, and the Worldwide Web

(WWW). Electronic mailing is the most useful resource extending the avenue of

learning to an international audience through news or list-servers groups.

Pediatric Surgery List Server, the most popular discussion group, is a constant

forum for exchange of ideas, difficult cases, consensus on management, and

development of our specialty. The WWW provides an all-in-one medium of text,

image, sound, and video. Associations, departments, educational sites,

organizations, peer-reviewed scientific journals and Medline database web pages

of prime interest to pediatric surgeons have been developing at an amazing pace.

Future developments of technological advance nurturing our specialty will consist

of online journals, telemedicine, international chatting, computer-based training

for surgical education, and centralization of cyberspace information into

database search sites.

PMID: 10688381 [Indexed for MEDLINE]

3715. Folia Neuropathol. 2000;38(1):43-6.

Real-time teleneuropathology for a second opinion of neurooncological cases.

Szymaś J(1), Papierz W, Danilewicz M.

Author information:

(1)Department of Pathology, Karol Marcinkowski School of Medicine, Poznań,

Poland. jszymas@ampat.amu.edu.pl

A teleneuropathology system directed by Java programs through a standard Internet

browser was evaluated in the present study. Assessment of neurooncological cases

was done by remote microscope through the "Case Study" Web page at the Department

of Pathology Web site (http:¿ampat.amu.edu.pl). The site was used to control a

remote automatic microscope, the Axioplan 2 (Zeiss), which was connected to a

computer that acts as an Internet server. The Java program for the control of the

microscope server is automatically downloaded and started if the user selects the

Web site of the corresponding microscope server. The microscope server receives

microscope operation commands from the telemicroscopy clients, executes them, and

distributes the new microscope image to all of the connected telemicroscopy

clients. Fifteen cases were evaluated over several weeks. The percentage of

correctly classified cases sent by remote consultation was 100%. Since the system

does not require specialized software for the remote side and since the number of

possible discussion partners is unlimited, this system may help to overcome

obstacles to the practice of teleneuropathology.

PMID: 11057034 [Indexed for MEDLINE]

3716. Nucleic Acids Res. 2000 Jan 1;28(1):304-5.

The ENZYME database in 2000.

Bairoch A(1).

Author information:

(1)Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel

Servet, 1211 Geneva 4, Switzerland. amos.bairoch@medecine.unige.ch

The ENZYME database is a repository of information related to the nomenclature of

enzymes. In recent years it has became an indispensable resource for the

development of metabolic databases. The current version contains information on

3705 enzymes. It is available through the ExPASy WWW server

(http://www.expasy.ch/enzyme/ ).

PMCID: PMC102465

PMID: 10592255 [Indexed for MEDLINE]

3717. Nucleic Acids Res. 2000 Jan 1;28(1):286-8.

The 1999 SWISS-2DPAGE database update.

Hoogland C(1), Sanchez JC, Tonella L, Binz PA, Bairoch A, Hochstrasser DF, Appel

RD.

Author information:

(1)Swiss Institute of Bioinformatics, 1 rue Michel-Servet, CH-1211 Genève 4,

Switzerland. christine.hoogland@isb-sib.ch

SWISS-2DPAGE (http://www.expasy.ch/ch2d/ ) is an annotated two-dimensional

polyacrylamide gel electro-phoresis (2-DE) database established in 1993. The

current release contains 24 reference maps from human and mouse biological

samples, as well as from Saccharomyces cerevisiae, Escherichia coli and

Dictyostelium discoideum origin. These reference maps have now 2824 identified

spots, corresponding to 614 separate protein entries in the database, in addition

to virtual entries for each SWISS-PROT sequence or any user-entered amino acids

sequence. Last year improvements in the SWISS-2DPAGE database are as follows:

three new maps have been created and several others have been updated;

cross-references to newly built federated 2-DE databases have been added; new

functions to access the data have been provided through the ExPASy proteomics

server.

PMCID: PMC102456

PMID: 10592248 [Indexed for MEDLINE]

3718. Nucleic Acids Res. 2000 Jan 1;28(1):277-82.

Assigning genomic sequences to CATH.

Pearl FM(1), Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM,

Orengo CA.

Author information:

(1)Department of Biochemistry, University College London, University of London,

Gower Street, London WC1E 6BT, UK. frances@biochem.ucl.ac.uk

We report the latest release (version 1.6) of the CATH protein domains database

(http://www.biochem.ucl. ac.uk/bsm/cath ). This is a hierarchical classification

of 18 577 domains into evolutionary families and structural groupings. We have

identified 1028 homo-logous superfamilies in which the proteins have both

structural, and sequence or functional similarity. These can be further clustered

into 672 fold groups and 35 distinct architectures. Recent developments of the

database include the generation of 3D templates for recognising structural

relatives in each fold group, which has led to significant improvements in the

speed and accuracy of updating the database and also means that less manual

validation is required. We also report the establishment of the CATH-PFDB

(Protein Family Database), which associates 1D sequences with the 3D homologous

superfamilies. Sequences showing identifiable homology to entries in CATH have

been extracted from GenBank using PSI-BLAST. A CATH-PSIBLAST server has been

established, which allows you to scan a new sequence against the database. The

CATH Dictionary of Homologous Superfamilies (DHS), which contains validated

multiple structural alignments annotated with consensus functional information

for evolutionary protein superfamilies, has been updated to include annotations

associated with sequence relatives identified in GenBank. The DHS is a powerful

tool for considering the variation of functional properties within a given CATH

superfamily and in deciding what functional properties may be reliably inherited

by a newly identified relative.

PMCID: PMC102424

PMID: 10592246 [Indexed for MEDLINE]

3719. Nucleic Acids Res. 2000 Jan 1;28(1):273-6.

ProClass protein family database.

Huang H(1), Xiao C, Wu CH.

Author information:

(1)Protein Information Resource, National Biomedical Research Foundation, 3900

Reservoir Road, NW, Washington, DC 20007, USA.

ProClass is a protein family database that organizes non-redundant sequence

entries into families defined collectively by PIR superfamilies and PROSITE

patterns. By combining global similarities and functional motifs into a single

classification scheme, ProClass helps to reveal domain and family relationships

and classify multi-domain proteins. The database currently consists of >155 000

sequence entries retrieved from both PIR-International and SWISS-PROT databases.

Approximately 92 000 or 60% of the ProClass entries are classified into

approximately 6000 families, including a large number of new members detected by

our GeneFIND family identification system. The ProClass motif collection contains

approximately 72 000 motif sequences and >1300 multiple alignments for all

PROSITE patterns, including >21 000 matches not listed in PROSITE and mostly

detected from unique PIR sequences. To maximize family information retrieval, the

database provides links to various protein family, domain, alignment and

structural class databases. With its high classification rate and comprehensive

family relationships, ProClass can be used to support full-scale genomic

annotation. The database, now being implemented in an object-relational database

management system, is available for online sequence search and record retrieval

from our WWW server at http://pir.georgetown.edu/gfserver/proclass.html

PMCID: PMC102450

PMID: 10592245 [Indexed for MEDLINE]

3720. Nucleic Acids Res. 2000 Jan 1;28(1):267-9.

ProDom and ProDom-CG: tools for protein domain analysis and whole genome

comparisons.

Corpet F(1), Servant F, Gouzy J, Kahn D.

Author information:

(1)Laboratoire de Génétique Cellulaire, INRA/CNRS, BP 27, F-31326

Castanet-Tolosan, Cedex, France.

ProDom contains all protein domain families automatically generated from the

SWISS-PROT and TrEMBL sequence databases (http://www.

toulouse.inra.fr/prodom.html ). ProDom-CG results from a similar domain analysis

as applied to completed genomes (http://www.toulouse. inra.fr/prodomCG.html ).

Recent improvements to the ProDom database and its server include: scaling up to

include sequences from TrEMBL, addition of Pfam-A entries to the set of expert

validated families, assignment of stable accession numbers, consistency

indicators for domain families, domain arrangements of sub-families and links to

Pfam-A.

PMCID: PMC102458

PMID: 10592243 [Indexed for MEDLINE]

3721. Nucleic Acids Res. 2000 Jan 1;28(1):219-21.

IMGT, the international ImMunoGeneTics database.

Ruiz M(1), Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, Marsh SG,

Bontrop R, Lemaitre M, Lefranc G, Chaume D, Lefranc MP.

Author information:

(1)Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UPR CNRS 1142 IGH, 141 rue de

la Cardonille, 34396 Montpellier Cedex 5, France.

IMGT, the international ImMunoGeneTics database (http://imgt.cines. fr:8104 ), is

a high-quality integrated database specialising in Immunoglobulins (Ig), T cell

Receptors (TcR) and Major Histocompatibility Complex (MHC) molecules of all

vertebrate species, created in 1989 by Marie-Paule Lefranc, Université

Montpellier II, CNRS, Montpellier, France (lefranc@ligm.igh.cnrs.fr ). At

present, IMGT includes two databases: IMGT/LIGM-DB, a comprehensive database of

Ig and TcR from human and other vertebrates, with translation for fully annotated

sequences, and IMGT/HLA-DB, a database of the human MHC referred to as HLA (Human

Leucocyte Antigens). The IMGT server provides a common access to expertized

genomic, proteomic, structural and polymorphic data of Ig and TcR molecules of

all vertebrates. By its high quality and its easy data distribution, IMGT has

important implications in medical research (repertoire in autoimmune diseases,

AIDS, leukemias, lymphomas), therapeutic approaches (antibody engineering),

genome diversity and genome evolution studies. IMGT is freely available at

http://imgt.cines.fr:8104. The IMGT Index is provided at the IMGT Marie-Paule

page (http://imgt.cines.fr:8104/textes/IMGTindex.html).

PMCID: PMC102442

PMID: 10592230 [Indexed for MEDLINE]

3722. Nucleic Acids Res. 2000 Jan 1;28(1):177-8.

The European large subunit ribosomal RNA database.

De Rijk P(1), Wuyts J, Van de Peer Y, Winkelmans T, De Wachter R.

Author information:

(1)Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1,

B-2610 Antwerpen, Belgium.

The European Large Subunit (LSU) Ribosomal RNA (rRNA) database is accessible via

the rRNA WWW Server at URL http://rrna.uia.ac.be/lsu/. It is a curated database

that compiles complete or nearly complete LSU rRNA sequences in aligned form, and

also incorporates secondary structure information for each sequence. Taxonomic

information, literature references and other information about the sequences are

also available, and can be searched via the WWW interface.

PMCID: PMC102430

PMID: 10592218 [Indexed for MEDLINE]

3723. Nucleic Acids Res. 2000 Jan 1;28(1):173-4.

The RDP (Ribosomal Database Project) continues.

Maidak BL(1), Cole JR, Lilburn TG, Parker CT Jr, Saxman PR, Stredwick JM, Garrity

GM, Li B, Olsen GJ, Pramanik S, Schmidt TM, Tiedje JM.

Author information:

(1)Center for Microbial Ecology, 540 Plant and Soil Sciences Building, Michigan

State University, East Lansing, MI 48824-1325, USA. curator@cme.msu.edu

The Ribosomal Database Project (RDP-II), previously described by Maidak et al.,

continued during the past year to add new rRNA sequences to the aligned data and

to improve the analysis commands. Release 7.1 (September 17, 1999) included more

than 10 700 small subunit rRNA sequences. More than 850 type strain sequences

were identified and added to the prokaryotic alignment, bringing the total number

of type sequences to 3324 representing 2460 different species. Availability of an

RDP-II mirror site in Japan is also near completion. RDP-II provides aligned and

annotated rRNA sequences, derived phylogenetic trees and taxonomic hierarchies,

and analysis services through its WWW server (http://rdp.cme.msu.edu/ ). Analysis

services include rRNA probe checking, approx-i-mate phylogenetic placement of

user sequences, screening user sequences for possible chimeric rRNA sequences,

automated alignment, production of similarity matrices and services to plan and

analyze terminal restriction fragment length polymorphism (T-RFLP) experiments.

PMCID: PMC102428

PMID: 10592216 [Indexed for MEDLINE]

3724. Nucleic Acids Res. 2000 Jan 1;28(1):68-71.

EMGLib: the enhanced microbial genomes library (update 2000).

Perrière G(1), Bessières P, Labedan B.

Author information:

(1)Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard,

Lyon 1, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France.

perriere@biomserv.univ-lyon.fr

As the number of complete microbial genomes publicly available is still growing,

the problem of annotation quality in these very large sequences remains unsolved.

Indeed, the number of annotations associated with complete genomes is usually

lower than those of the shorter entries encountered in the repository

collections. Moreover, classical sequence database management systems have

difficulties in handling entries of such size. In this context, the Enhanced

Microbial Genomes Library (EMGLib) was developed to try to alleviate these

problems. This library contains all the complete genomes from prokaryotes

(bacteria and archaea) already sequenced and the yeast genome in GenBank format.

The annotations are improved by the introduction of data on codon usage, gene

orientation on the chromosome and gene families. It is possible to access EMGLib

through two database systems set up on WWW servers: the PBIL server at

http://pbil.univ-lyon1.fr/emglib.html and the MICADO server at

http://locus.jouy.inra.fr/micado

PMCID: PMC102414

PMID: 10592183 [Indexed for MEDLINE]

3725. Nucleic Acids Res. 2000 Jan 1;28(1):49-55.

ProtoMap: automatic classification of protein sequences and hierarchy of protein

families.

Yona G(1), Linial N, Linial M.

Author information:

(1)Department of Structural Biology, Fairchild Building D-109, Stanford

University, CA 94305, USA. golan@gimmel.stanford.edu

The ProtoMap site offers an exhaustive classification of all proteins in the

SWISS-PROT database, into groups of related proteins. The classification is based

on analysis of all pairwise similarities among protein sequences. The analysis

makes essential use of transitivity to identify homologies among proteins. Within

each group of the classification, every two members are either directly or

transitively related. However, transitivity is applied restrictively in order to

prevent unrelated proteins from clustering together. The classification is done

at different levels of confidence, and yields a hierarchical organization of all

proteins. The resulting classification splits the protein space into well-defined

groups of proteins, which are closely correlated with natural biological families

and superfamilies. Many clusters contain protein sequences that are not

classified by other databases. The hierarchical organization suggested by our

analysis may help in detecting finer subfamilies in families of known proteins.

In addition it brings forth interesting relationships between protein families,

upon which local maps for the neighborhood of protein families can be sketched.

The ProtoMap web server can be accessed at http://www.protomap.cs.huji.ac.il

PMCID: PMC102438

PMID: 10592179 [Indexed for MEDLINE]

3726. Nucleic Acids Res. 2000 Jan 1;28(1):37-40.

MIPS: a database for genomes and protein sequences.

Mewes HW(1), Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt

G, Pfeiffer F, Schüller C, Stocker S, Weil B.

Author information:

(1)GSF-Forschungszentrum für Umwelt und Gesundheit, Munich Information Center for

Protein Sequences, am Max-Planck-Institut für Biochemie, Am Klopferspitz 18,

D-82152 Martinsried, Germany. mewes@mips.embuct.org

The Munich Information Center for Protein Sequences (MIPS-GSF), Martinsried, near

Munich, Germany, continues its longstanding tradition to develop and maintain

high quality curated genome databases. In addition, efforts have been intensified

to cover the wealth of complete genome sequences in a systematic, comprehensive

form. Bioinformatics, supporting national as well as European sequencing and

functional analysis projects, has resulted in several up-to-date genome-oriented

databases. This report describes growing databases reflecting the progress of

sequencing the Arabidopsis thaliana (MATDB) and Neurospora crassa genomes

(MNCDB), the yeast genome database (MYGD) extended by functional analysis data,

the database of annotated human EST-clusters (HIB) and the database of the

complete cDNA sequences from the DHGP (German Human Genome Project). It also

contains information on the up-to-date database of complete genomes (PEDANT), the

classification of protein sequences (ProtFam) and the collection of protein

sequence data within the framework of the PIR-International Protein Sequence

Database. These databases can be accessed through the MIPS WWW server

(http://www. mips.biochem.mpg.de).

PMCID: PMC102494

PMID: 10592176 [Indexed for MEDLINE]

3727. Nucleic Acids Res. 2000 Jan 1;28(1):8-9.

DBcat: a catalog of 500 biological databases.

Discala C(1), Benigni X, Barillot E, Vaysseix G.

Author information:

(1)CNS, 2 rue Gaston Crémieux, BP 191, 91006 Evry cedex, France.

The DBcat (http://www.infobiogen.fr/services/dbcat ) is a comprehensive catalog

of biological databases, maintained and curated at Infobiogen. It contains 500

databases classified by application domains. The DBcat is a structured flat-file

library, that can be searched by means of an SRS server or a dedicated Web

interface. The files are available for download from Infobiogen anonymous ftp

server.

PMCID: PMC102454

PMID: 10592168 [Indexed for MEDLINE]

3728. Stud Health Technol Inform. 2000;78:101-25.

Sleep atlas and multimedia database.

Penzel T(1), Kesper K, Mayer G, Zulley J, Peter JH.

Author information:

(1)Sleep Laboratory, Department of Internal Medicine, Philipps-University,

Baldingerstr. 1, D-35033 Marburg, Germany. penzel@Mailer.uni-marburg.de

The ENN sleep atlas and database was set up on a dedicated server connected to

the internet thus providing all services such as WWW, ftp and telnet access. The

database serves as a platform to promote the goals of the European Neurological

Network, to exchange patient cases for second opinion between experts and to

create a case-oriented multimedia sleep atlas with descriptive text, images and

video-clips of all known sleep disorders. The sleep atlas consists of a small

public and a large private part for members of the consortium. 20 patient cases

were collected and presented with educational information similar to published

case reports. Case reports are complemented with images, video-clips and

biosignal recordings. A Java based viewer for biosignals provided in EDF format

was installed in order to move free within the sleep recordings without the need

to download the full recording on the client.

PMID: 11151592 [Indexed for MEDLINE]

3729. J Mol Biol. 1999 Dec 17;294(5):1351-62.

Sequence and structure-based prediction of eukaryotic protein phosphorylation

sites.

Blom N(1), Gammeltoft S, Brunak S.

Author information:

(1)Department of Biotechnology, The Technical University of Denmark, Lyngby,

DK-2800, Denmark.

Protein phosphorylation at serine, threonine or tyrosine residues affects a

multitude of cellular signaling processes. How is specificity in substrate

recognition and phosphorylation by protein kinases achieved? Here, we present an

artificial neural network method that predicts phosphorylation sites in

independent sequences with a sensitivity in the range from 69 % to 96 %. As an

example, we predict novel phosphorylation sites in the p300/CBP protein that may

regulate interaction with transcription factors and histone acetyltransferase

activity. In addition, serine and threonine residues in p300/CBP that can be

modified by O-linked glycosylation with N-acetylglucosamine are identified.

Glycosylation may prevent phosphorylation at these sites, a mechanism named

yin-yang regulation. The prediction server is available on the Internet at

http://www.cbs.dtu.dk/services/NetPhos/or via e-mail to NetPhos@cbs. dtu.dk.

Copyright 1999 Academic Press.

DOI: 10.1006/jmbi.1999.3310

PMID: 10600390 [Indexed for MEDLINE]

3730. J Mol Biol. 1999 Dec 10;294(4):921-35.

kPROT: a knowledge-based scale for the propensity of residue orientation in

transmembrane segments. Application to membrane protein structure prediction.

Pilpel Y(1), Ben-Tal N, Lancet D.

Author information:

(1)Department of Molecular Genetics and the Crown Genome Center, The Weizmann

Institute of Science, Rehovot, 76100, Israel. bnpilpel@membran1.weizmann.ac.il

Modeling of integral membrane proteins and the prediction of their functional

sites requires the identification of transmembrane (TM) segments and the

determination of their angular orientations. Hydrophobicity scales predict

accurately the location of TM helices, but are less accurate in computing angular

disposition. Estimating lipid-exposure propensities of the residues from

statistics of solved membrane protein structures has the disadvantage of relying

on relatively few proteins. As an alternative, we propose here a scale of

knowledge-based Propensities for Residue Orientation in Transmembrane segments

(kPROT), derived from the analysis of more than 5000 non-redundant protein

sequences. We assume that residues that tend to be exposed to the membrane are

more frequent in TM segments of single-span proteins, while residues that prefer

to be buried in the transmembrane bundle interior are present mainly in

multi-span TMs. The kPROT value for each residue is thus defined as the logarithm

of the ratio of its proportions in single and multiple TM spans. The scale is

refined further by defining it for three discrete sections of the TM segment;

namely, extracellular, central, and intracellular. The capacity of the kPROT

scale to predict angular helical orientation was compared to that of alternative

methods in a benchmark test, using a diversity of multi-span alpha-helical

transmembrane proteins with a solved 3D structure. kPROT yielded an average

angular error of 41 degrees, significantly lower than that of alternative scales

(62 degrees -68 degrees ). The new scale thus provides a useful general tool for

modeling and prediction of functional residues in membrane proteins. A WWW server

(http://bioinfo.weizmann.ac.il/kPROT) is available for automatic helix

orientation prediction with kPROT.

Copyright 1999 Academic Press.

DOI: 10.1006/jmbi.1999.3257

PMID: 10588897 [Indexed for MEDLINE]

3731. Eur J Surg. 1999 Dec;165(12):1121-4.

Lessons from analysis of World Wide Web server activity data:

http://www.swsahs.nsw.gov.au/livtrauma.

Ryan J(1), Sugrue M, Geller E, Lu W, Kolkman K.

Author information:

(1)Department of Trauma Services, Liverpool Hospital, NSW, Australia.

OBJECTIVE: To establish visiting patterns and use of resources within an

educational world wide web site, to develop strategies for its more effective use

as a teaching medium.

DESIGN: Prospective descriptive study.

SETTING: A web site maintained by a major Australian metropolitan trauma service.

SUBJECTS: All visitors to the web site over a 40 day period.

MAIN OUTCOME MEASURES: Number of visitors, requests for pages, and transfer of

files; timing and duration of visits; first and last pages viewed during visits;

pages viewed most often; source of referral and country of origin of visitors.

RESULTS: There were 2237 visits, and a mean of 5.8 pages/visit were viewed. Mean

duration of each visit was 4 minutes 35 seconds, and only half the visitors

entered through the home page. The collection of radiographs was most commonly

consulted. Visitors came from 26 different countries, 42% from Australia.

CONCLUSIONS: Factors that influenced the use of the site were identified and have

altered our plans for development.

DOI: 10.1080/110241599750007612

PMID: 10636542 [Indexed for MEDLINE]

3732. J Comput Aided Mol Des. 1999 Nov;13(6):625-43.

Visualisation and integration of G protein-coupled receptor related information

help the modelling: description and applications of the Viseur program.

Campagne F(1), Jestin R, Reversat JL, Bernassau JM, Maigret B.

Author information:

(1)Laboratoire de Chimie théorique, Centre Charles Hermite, CNRS-Université Henri

Poincaré Nancy I, Vandoeuvre-les-Nancy, France. campagne@inka.mssm.edu

G Protein-Coupled Receptors (GPCRs) constitute a superfamily of receptors that

forms an important therapeutic target. The number of known GPCR sequences and

related information increases rapidly. For these reasons, we are developing the

Viseur program to integrate the available information related to GPCRs. The

Viseur program allows one to interactively visualise and/or modify the sequences,

transmembrane areas, alignments, models and results of mutagenesis experiments in

an integrated environment. This integration increases the ease of modelling

GPCRs: visualisation and manipulation improvements enable easier databank

interrogation and interpretation. Unique program features include: (i) automatic

construction of 'Snake-like' diagrams or hyperlinked GPCR molecular models to

HTML or VRML and (ii) automatic access to a mutagenesis data server through the

Internet. The novel algorithms or methods involved are presented, followed by the

overall complementary features of the program. Finally, we present two

applications of the program: (i) an automatic construction of GPCR snake-like

diagrams for the GPCRDB WWW server, and (ii) a preparation of the modelling of

the 5HT receptor subtypes. The interest of the direct access to mutagenesis

results through an alignment and a molecular model are discussed. The Viseur

program, which runs on SGI workstations, is freely available and can be used for

preparing the modelling of integral membrane proteins or as an alignment editor

tool.

PMID: 10584220 [Indexed for MEDLINE]

3733. Glycobiology. 1999 Oct;9(10):1009-22.

Scanning the available Dictyostelium discoideum proteome for O-linked GlcNAc

glycosylation sites using neural networks.

Gupta R(1), Jung E, Gooley AA, Williams KL, Brunak S, Hansen J.

Author information:

(1)Department of Biotechnology, Technical University of Denmark, Lyngby, Denmark.

Dictyostelium discoideum has been suggested as a eukaryotic model organism for

glycobiology studies. Presently, the characteristics of acceptor sites for the

N-acetylglucosaminyl-transferases in Dictyostelium discoideum, which link GlcNAc

in an alpha linkage to hydroxyl residues, are largely unknown. This motivates the

development of a species specific method for prediction of O-linked GlcNAc

glycosylation sites in secreted and membrane proteins of D. discoideum. The

method presented here employs a jury of artificial neural networks. These

networks were trained to recognize the sequence context and protein surface

accessibility in 39 experimentally determined O-alpha-GlcNAc sites found in D.

discoideum glycoproteins expressed in vivo. Cross-validation of the data revealed

a correlation in which 97% of the glycosylated and nonglycosylated sites were

correctly identified. Based on the currently limited data set, an abundant

periodicity of two (positions-3, -1, +1, +3, etc.) in Proline residues

alternating with hydroxyl amino acids was observed upstream and downstream of the

acceptor site. This was a consequence of the spacing of the glycosylated residues

themselves which were peculiarly found to be situated only at even positions with

respect to each other, indicating that these may be located within beta-strands.

The method has been used for a rapid and ranked scan of the fraction of the

Dictyostelium proteome available in public databases, remarkably 25-30% of which

were predicted glycosylated. The scan revealed acceptor sites in several proteins

known experimentally to be O-glycosylated at unmapped sites. The available

proteome was classified into functional and cellular compartments to study any

preferential patterns of glycosylation. A sequence based prediction server for

GlcNAc O-glycosylations in D. discoideum proteins has been made available through

the WWW at http://www.cbs.dtu.dk/services/DictyOGlyc/ and via E-mail to

DictyOGlyc@cbs.dtu.dk.

PMID: 10521537 [Indexed for MEDLINE]

3734. Radiat Meas. 1999 Oct;30(5):653-9.

Computer simulation of spacecraft/environment interaction.

Krupnikov KK(1), Makletsov AA, Mileev VN, Novikov LS, Sinolits VV.

Author information:

(1)Skobeltsyn Institute of Nuclear Physics, Moscow State University, Russia.

This report presents some examples of a computer simulation of spacecraft

interaction with space environment. We analysed a set data on electron and ion

fluxes measured in 1991 1994 on geostationary satellite GORIZONT-35. The

influence of spacecraft eclipse and device eclipse by solar-cell panel on

spacecraft charging was investigated. A simple method was developed for an

estimation of spacecraft potentials in LEO. Effects of various particle flux

impact and spacecraft orientation are discussed. A computer engineering model for

a calculation of space radiation is presented. This model is used as a

client/server model with WWW interface, including spacecraft model description

and results representation based on the virtual reality markup language.

PMID: 11542669 [Indexed for MEDLINE]

3735. Biofizika. 1999 Sep-Oct;44(5):832-6.

[WWWMGS: an integrated server for molecular-genetic studies].

[Article in Russian]

Frolov AS(1), Lavriushev SV, Grigorovich DA, Kel AE, Ptitsyn AA, Kolchanov NA,

Podkolodnyĭ NL, Solov'ev VV, Milanesi L, Bourne P, et al.

Author information:

(1)Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk,

Russia.

We report an integrative technology for molecular biology studies in the field of

transcription regulation by using Internet. A set of databases, programs, and

systems are included into WWWMGS Web server. For example, the use of TRRD

database information for site prediction is described. Using this method, the

computer system SeqAnn was developed. The system performs the "real time"

searching for prediction of initiation transcription site position according to

database information. WWWMGS is available at URL: http://wwwmgs.bionet.nsc.ru/.

PMID: 10624522 [Indexed for MEDLINE]

3736. Electrophoresis. 1999 Aug;20(11):2280-98.

Two-dimensional gel protein database of Saccharomyces cerevisiae (update 1999).

Perrot M(1), Sagliocco F, Mini T, Monribot C, Schneider U, Shevchenko A, Mann M,

Jenö P, Boucherie H.

Author information:

(1)Institut de Biochimie et Génétique Cellulaires, UPR CNRS 9026, Bordeaux,

France.

By proving the opportunity to visualize several hundred proteins at a time,

two-dimensional (2-D) gel electrophoresis is an important tool for proteome

research. In order to take advantage of the full potential of this technique for

yeast studies, we have undertaken a systematic identification of yeast proteins

resolved by this technique. We report here the identification of 92 novel protein

spots on the yeast 2-D protein map. These identifications extend the number of

protein spots identified on our yeast reference map to 401. These spots

correspond to the products of 279 different genes. They have been essentially

identified by three methods: gene overexpression, amino acid composition and mass

spectrometry. These data can be accessed on the Yeast Protein Map server

(htpp://www.ibgc.u-bordeaux2.fr/YPM).

DOI: 10.1002/(SICI)1522-2683(19990801)20:11<2280::AID-ELPS2280>3.0.CO;2-Q

PMID: 10493132 [Indexed for MEDLINE]

3737. Protein Eng. 1999 Aug;12(8):631-4.

An hierarchical artificial neural network system for the classification of

transmembrane proteins.

Pasquier C(1), Hamodrakas SJ.

Author information:

(1)Faculty of Biology, Department of Cell Biology and Biophysics, University of

Athens, Panepistimiopolis, Athens 15701, Greece.

This work presents a simple artificial neural network which classifies proteins

into two classes from their sequences alone: the membrane protein class and the

non-membrane protein class. This may be important in the functional assignment

and analysis of open reading frames (ORF's) identified in complete genomes and,

especially, those ORF's that correspond to proteins with unknown function. The

network described here has a simple hierarchical feed-forward topology and a

limited number of neurons which makes it very fast. By using only information

contained in 11 protein sequences, the method was able to identify, with 100%

accuracy, all membrane proteins with reliable topologies collected from several

papers in the literature. Applied to a test set of 995 globular, water-soluble

proteins, the neural network classified falsely 23 of them in the membrane

protein class (97.7% of correct assignment). The method was also applied to the

complete SWISS-PROT database with considerable success and on ORF's of several

complete genomes. The neural network developed was associated with the PRED-TMR

algorithm (Pasquier,C., Promponas,V.J., Palaios,G.A., Hamodrakas,J.S. and

Hamodrakas,S.J., 1999) in a new application package called PRED-TMR2. A WWW

server running the PRED-TMR2 software is available at

http://o2.db.uoa.gr/PRED-TMR2

PMID: 10469822 [Indexed for MEDLINE]

3738. Bioinformatics. 1999 Jun;15(6):523-4.

BLAST PRINTS--alternative perspectives on sequence similarity.

Wright W(1), Scordis P, Attwood TK.

Author information:

(1)School of Biological Sciences, Stopford Building, The University of

Manchester, Oxford Road, Manchester M13 6PT, UK. wright@biochemistry.ucl.ac.uk

SUMMARY: An implementation of BLAST for searching the PRINTS database is

presented. The interface allows submission of either protein or DNA queries, and

returns the familiar form of output, but modified by means of direct links both

to the familial discriminators in PRINTS and to fingerprint profile visualization

software. The server thus couples the rapidity of BLAST searching with the

sensitivity of fingerprint diagnoses, providing alternative perspectives on a

given query.

AVAILABILITY: http://www.biochem.ucl. ac.uk/cgi-bin/wright/printsBLAST.cgi

PMID: 10383477 [Indexed for MEDLINE]

3739. Bioinformatics. 1999 Jun;15(6):510-20.

An ontology for bioinformatics applications.

Baker PG(1), Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A.

Author information:

(1)School of Biological Sciences and Department of Computer Science, University

of Manchester, Oxford Road, Manchester M13 9PT, UK. tambis@cs.man.ac.uk

MOTIVATION: An ontology of biological terminology provides a model of biological

concepts that can be used to form a semantic framework for many data storage,

retrieval and analysis tasks. Such a semantic framework could be used to underpin

a range of important bioinformatics tasks, such as the querying of heterogeneous

bioinformatics sources or the systematic annotation of experimental results.

RESULTS: This paper provides an overview of an ontology [the Transparent Access

to Multiple Biological Information Sources (TAMBIS) ontology or TaO] that

describes a wide range of bioinformatics concepts. The present paper describes

the mechanisms used for delivering the ontology and discusses the ontology's

design and organization, which are crucial for maintaining the coherence of a

large collection of concepts and their relationships.

AVAILABILITY: The TAMBIS system, which uses a subset of the TaO described here,

is accessible over the Web via http://img.cs.man.ac.uk/tambis (although in the

first instance, we will use a password mechanism to limit the load on our

server). The complete model is also available on the Web at the above URL.

PMID: 10383475 [Indexed for MEDLINE]

3740. Genome Res. 1999 Jun;9(6):AP1-8, insert.

A high-density integrated genetic linkage and radiation hybrid map of the

laboratory rat.

Steen RG(1), Kwitek-Black AE, Glenn C, Gullings-Handley J, Van Etten W, Atkinson

OS, Appel D, Twigger S, Muir M, Mull T, Granados M, Kissebah M, Russo K, Crane R,

Popp M, Peden M, Matise T, Brown DM, Lu J, Kingsmore S, Tonellato PJ, Rozen S,

Slonim D, Young P, Jacob HJ.

Author information:

(1)Center for Genome Research, Whitehead Institute for Biomedical Research and

Massachusetts Institute of Technology, Cambridge, Massachusetts 02142 USA.

Erratum in

Genome Res 1999 Aug;9(8):793.

The laboratory rat (Rattus norvegicus) is a key animal model for biomedical

research. However, the genetic infrastructure required for connecting phenotype

and genotype in the rat is currently incomplete. Here, we report the construction

and integration of two genomic maps: a dense genetic linkage map of the rat and

the first radiation hybrid (RH) map of the rat. The genetic map was constructed

in two F2 intercrosses (SHRSP x BN and FHH x ACI), containing a total of 4736

simple sequence length polymorphism (SSLP) markers. Allele sizes for 4328 of the

genetic markers were characterized in 48 of the most commonly used inbred

strains. The RH map is a lod >/= 3 framework map, including 983 SSLPs, thereby

allowing integration with markers on various genetic maps and with markers mapped

on the RH panel. Together, the maps provide an integrated reference to >3000

genes and ESTs and >8500 genetic markers (5211 of our SSLPs and >3500 SSLPs

developed by other groups). [Bihoreau et al. (1997); James and Tanigami, RHdb

(http:www.ebi.ac.uk/RHdb/index.html); Wilder

(http://www.nih.gov/niams/scientific/ratgbase); Serikawa et al. (1992); RATMAP

server (http://ratmap.gen.gu.se)] RH maps (v. 2.0) have been posted on our web

sites at http://goliath.ifrc.mcw.edu/LGR/index.html or

http://curatools.curagen.com/ratmap. Both web sites provide an RH mapping server

where investigators can localize their own RH vectors relative to this map. The

raw data have been deposited in the RHdb database. Taken together, these maps

provide the basic tools for rat genomics. The RH map provides the means to

rapidly localize genetic markers, genes, and ESTs within the rat genome. These

maps provide the basic tools for rat genomics. They will facilitate studies of

multifactorial disease and functional genomics, allow construction of physical

maps, and provide a scaffold for both directed and large-scale sequencing efforts

and comparative genomics in this important experimental organism.

PMID: 10400928 [Indexed for MEDLINE]

3741. Chem Biol Interact. 1999 May 14;119-120:567-76.

Kinetic parameters of cholinesterase interactions with organophosphates:

retrieval and comparison tools available through ESTHER database: ESTerases,

alpha/beta Hydrolase Enzymes and Relatives.

Chatonnet A(1), Hotelier T, Cousin X.

Author information:

(1)Différenciation Cellulaire et Croissance, INRA-ENSAM, Montpellier, France.

chedb@ensam.inra.fr

Cholinesterases are targets for organophosphorus compounds which are used as

insecticides, chemical warfare agents and drugs for the treatment of disease such

as glaucoma, or parasitic infections. The widespread use of these chemicals

explains the growing of this area of research and the ever increasing number of

sequences, structures, or biochemical data available. Future advances will depend

upon effective management of existing information as well as upon creation of new

knowledge. The ESTHER database goal is to facilitate retrieval and comparison of

data about structure and function of proteins presenting the alpha/beta hydrolase

fold. Protein engineering and in vitro production of enzymes allow direct

comparison of biochemical parameters. Kinetic parameters of enzymatic reactions

are now included in the database. These parameters can be searched and compared

with a table construction tool. ESTHER can be reached through internet

(http://www.ensam.inra.fr/cholinesterase). The full database or the specialised

X-window Client-server system can be downloaded from our ftp server

(ftp://ftp.toulouse.inra.fr./pub/esther). Forms can be used to send updates or

corrections directly from the web.

PMID: 10421496 [Indexed for MEDLINE]

3742. Bioinformatics. 1999 May;15(5):426-7.

REPuter: fast computation of maximal repeats in complete genomes.

Kurtz S(1), Schleiermacher C.

Author information:

(1)Technische Fakultät, Universität Bielefeld, Postfach 10 01 31, D-33501

Bielefeld, Germany. kurtz.icschlei@techfak.uni-bielefeld.de

SUMMARY: A software tool was implemented that computes exact repeats and

palindromes in entire genomes very efficiently.

AVAILABILITY: Via the Bielefeld Bioinformatics Server

(http://bibiserv.techfak.uni-bielefeld.de/rep uter/).

PMID: 10366664 [Indexed for MEDLINE]

3743. Bioinformatics. 1999 May;15(5):413-21.

Improved performance in protein secondary structure prediction by inhomogeneous

score combination.

Guermeur Y(1), Geourjon C, Gallinari P, Deléage G.

Author information:

(1)LIP, Ecole Normale Supérieure de Lyon, 46, Allée d'Italie, 69364 Lyon cedex

07, France. Yann.Guermeur@ens-lyon.fr

MOTIVATION: In many fields of pattern recognition, combination has proved

efficient to increase the generalization performance of individual prediction

methods. Numerous systems have been developed for protein secondary structure

prediction, based on different principles. Finding better ensemble methods for

this task may thus become crucial. Furthermore, efforts need to be made to help

the biologist in the post-processing of the outputs.

RESULTS: An ensemble method has been designed to post-process the outputs of

discriminant models, in order to obtain an improvement in prediction accuracy

while generating class posterior probability estimates. Experimental results

establish that it can increase the recognition rate of protein secondary

structure prediction methods that provide inhomogeneous scores, even though their

individual prediction successes are largely different. This combination thus

constitutes a help for the biologist, who can use it confidently on top of any

set of prediction methods. Moreover, the resulting estimates can be used in

various ways, for instance to determine which areas in the sequence are predicted

with a given level of reliability.

AVAILABILITY: The prediction is freely available over the Internet on the Network

Protein Sequence Analysis (NPS@) WWW server at

http://pbil.ibcp.fr/NPSA/npsa\_server.ht ml. The source code of the combiner can

be obtained on request for academic use.

PMID: 10366661 [Indexed for MEDLINE]

3744. Bioinformatics. 1999 May;15(5):391-412.

Automated genome sequence analysis and annotation.

Andrade MA(1), Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A,

Tamames J, Valencia A, Ouzounis C, Sander C.

Author information:

(1)European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus,

Cambridge CB10 1SD, UK.

MOTIVATION: Large-scale genome projects generate a rapidly increasing number of

sequences, most of them biochemically uncharacterized. Research in bioinformatics

contributes to the development of methods for the computational characterization

of these sequences. However, the installation and application of these methods

require experience and are time consuming.

RESULTS: We present here an automatic system for preliminary functional

annotation of protein sequences that has been applied to the analysis of sets of

sequences from complete genomes, both to refine overall performance and to make

new discoveries comparable to those made by human experts. The GeneQuiz system

includes a Web-based browser that allows examination of the evidence leading to

an automatic annotation and offers additional information, views of the results,

and links to biological databases that complement the automatic analysis. System

structure and operating principles concerning the use of multiple sequence

databases, underlying sequence analysis tools, lexical analyses of database

annotations and decision criteria for functional assignments are detailed. The

system makes automatic quality assessments of results based on prior experience

with the underlying sequence analysis tools; overall error rates in functional

assignment are estimated at 2.5-5% for cases annotated with highest reliability

('clear' cases). Sources of over-interpretation of results are discussed with

proposals for improvement. A conservative definition for reporting 'new findings'

that takes account of database maturity is presented along with examples of

possible kinds of discoveries (new function, family and superfamily) made by the

system. System performance in relation to sequence database coverage, database

dynamics and database search methods is analysed, demonstrating the inherent

advantages of an integrated automatic approach using multiple databases and

search methods applied in an objective and repeatable manner.

AVAILABILITY: The GeneQuiz system is publicly available for analysis of protein

sequences through a Web server at http://www.sander.ebi.ac. uk/gqsrv/submit

PMID: 10366660 [Indexed for MEDLINE]

3745. Bioinformatics. 1999 May;15(5):356-61.

Promoter2.0: for the recognition of PolII promoter sequences.

Knudsen S(1).

Author information:

(1)Center for Biological Sequence Analysis, The Technical University of Denmark,

DK-2800 Lyngby, Denmark. steen@cbs.dtu.dk

MOTIVATION: A new approach to the prediction of eukaryotic PolII promoters from

DNA sequence takes advantage of a combination of elements similar to neural

networks and genetic algorithms to recognize a set of discrete subpatterns with

variable separation as one pattern: a promoter. The neural networks use as input

a small window of DNA sequence, as well as the output of other neural networks.

Through the use of genetic algorithms, the weights in the neural networks are

optimized to discriminate maximally between promoters and non-promoters.

RESULTS: After several thousand generations of optimization, the algorithm was

able to discriminate between vertebrate promoter and non-promoter sequences in a

test set with a correlation coefficient of 0.63. In addition, all five known

transcription start sites on the plus strand of the complete adenovirus genome

were within 161 bp of 35 predicted transcription start sites. On standardized

test sets consisting of human genomic DNA, the performance of Promoter2.0

compares well with other software developed for the same purpose.

AVAILABILITY: Promoter2.0 is available as a Web server at http://www.cbs.dtu.

dk/services/promoter/

PMID: 10366655 [Indexed for MEDLINE]

3746. J Digit Imaging. 1999 May;12(2 Suppl 1):205-7.

Interhospital network system using the worldwide web and the common gateway

interface.

Oka A(1), Harima Y, Nakano Y, Tanaka Y, Watanabe A, Kihara H, Sawada S.

Author information:

(1)Department of Radiology, Kansai Medical University, Osaka, Japan.

We constructed an interhospital network system using the worldwide web (WWW) and

the Common Gateway Interface (CGI). Original clinical images are digitized and

stored as a database for educational and research purposes. Personal computers

(PCs) are available for data treatment and browsing. Our system is simple, as

digitized images are stored into a Unix server machine. Images of important and

interesting clinical cases are selected and registered into the image database

using CGI. The main image format is 8- or 12-bit Joint Photographic Experts Group

(JPEG) image. Original clinical images are finally stored in CD-ROM using a CD

recorder. The image viewer can browse all of the images for one case at once as

thumbnail pictures; image quality can be selected depending on the user's

purpose. Using the network system, clinical images of interesting cases can be

rapidly transmitted and discussed with other related hospitals. Data transmission

from relational hospitals takes 1 to 2 minutes per 500 Kbyte of data. More

distant hospitals (e.g., Rakusai Hospital, Kyoto) takes 1 minute more. The mean

number of accesses our image database in a recent 3-month period was 470. There

is a total about 200 cases in our image database, acquired over the past 2 years.

Our system is useful for communication and image treatment between hospitals and

we will describe the elements of our system and image database.

PMCID: PMC3452880

PMID: 10342215 [Indexed for MEDLINE]

3747. J Digit Imaging. 1999 May;12(2 Suppl 1):175-7.

Transparent image access in a distributed picture archiving and communications

system: the Master Database broker.

Cox RD(1), Henri CJ, Rubin RK.

Author information:

(1)Department of Diagnostic Radiology, McGill University Health Centre, Montreal,

Quebec, Canada.

A distributed design is the most cost-effective system for small-to medium-scale

picture archiving and communications systems (PACS) implementations. However, the

design presents an interesting challenge to developers and implementers: to make

stored image data, distributed throughout the PACS network, appear to be

centralized with a single access point for users. A key component for the

distributed system is a central or master database, containing all the studies

that have been scanned into the PACS. Each study includes a list of one or more

locations for that particular dataset so that applications can easily find it.

Non-Digital Imaging and Communications in Medicine (DICOM) clients, such as our

worldwide web (WWW)-based PACS browser, query the master database directly to

find the images, then jump to the most appropriate location via a distributed

web-based viewing system. The Master Database Broker provides DICOM clients with

the same functionality by translating DICOM queries to master database searches

and distributing retrieval requests transparently to the appropriate source. The

Broker also acts as a storage service class provider, allowing users to store

selected image subsets and reformatted images with the original study, without

having to know on which server the original data are stored.

PMCID: PMC3452888

PMID: 10342203 [Indexed for MEDLINE]

3748. Protein Eng. 1999 May;12(5):381-5.

A novel method for predicting transmembrane segments in proteins based on a

statistical analysis of the SwissProt database: the PRED-TMR algorithm.

Pasquier C(1), Promponas VJ, Palaios GA, Hamodrakas JS, Hamodrakas SJ.

Author information:

(1)Faculty of Biology, Department of Cell Biology and Biophysics, University of

Athens, Panepistimiopolis, Athens 15701, Greece.

We present a novel method that predicts transmembrane domains in proteins using

solely information contained in the sequence itself. The PRED-TMR algorithm

described, refines a standard hydrophobicity analysis with a detection of

potential termini ('edges', starts and ends) of transmembrane regions. This

allows one both to discard highly hydrophobic regions not delimited by clear

start and end configurations and to confirm putative transmembrane segments not

distinguishable by their hydrophobic composition. The accuracy obtained on a test

set of 101 non-homologous transmembrane proteins with reliable topologies

compares well with that of other popular existing methods. Only a slight decrease

in prediction accuracy was observed when the algorithm was applied to all

transmembrane proteins of the SwissProt database (release 35). A WWW server

running the PRED-TMR algorithm is available at http://o2.db.uoa. gr/PRED-TMR/

PMID: 10360978 [Indexed for MEDLINE]

3749. Protein Sci. 1999 May;8(5):978-84.

ChloroP, a neural network-based method for predicting chloroplast transit

peptides and their cleavage sites.

Emanuelsson O(1), Nielsen H, von Heijne G.

Author information:

(1)Department of Biochemistry, Stockholm University, Sweden.

We present a neural network based method (ChloroP) for identifying chloroplast

transit peptides and their cleavage sites. Using cross-validation, 88% of the

sequences in our homology reduced training set were correctly classified as

transit peptides or nontransit peptides. This performance level is well above

that of the publicly available chloroplast localization predictor PSORT. Cleavage

sites are predicted using a scoring matrix derived by an automatic motif-finding

algorithm. Approximately 60% of the known cleavage sites in our sequence

collection were predicted to within +/-2 residues from the cleavage sites given

in SWISS-PROT. An analysis of 715 Arabidopsis thaliana sequences from SWISS-PROT

suggests that the ChloroP method should be useful for the identification of

putative transit peptides in genome-wide sequence data. The ChloroP predictor is

available as a web-server at http://www.cbs.dtu.dk/services/ChloroP/.

DOI: 10.1110/ps.8.5.978

PMCID: PMC2144330

PMID: 10338008 [Indexed for MEDLINE]

3750. Anasthesiol Intensivmed Notfallmed Schmerzther. 1999 Apr;34(4):251-3.

[New on the Internet: www.thieme.de/ains. The theme server: www.anaesthesie.com].

[Article in German]

Lutz JF(1).

Author information:

(1)Klinik für Anästhesiologie und operative Intensivmedizin Katharinenhospital,

Stuttgart.

DOI: 10.1055/s-1999-180

PMID: 10352808 [Indexed for MEDLINE]

3751. Bioinformatics. 1999 Apr;15(4):343-4.

The domain-server: direct prediction of protein domain-homologies from BLAST

search.

Murvai J(1), Vlahovicek K, Barta E, Parthasarathy S, Hegyi H, Pfeiffer F, Pongor

S.

Author information:

(1)International Centre for Genetic Engineering and Biotechnology, 34012 Trieste,

Italy.

RESULTS: A WWW server for protein domain homology prediction, based on BLAST

search and a simple data-mining algorithm (Hegyi,H. and Pongor,S. (1993) Comput.

Appl. Biosci., 9, 371-372), was constructed providing a tabulated list and a

graphic plot of similarities.

AVAILABILITY: http://www.icgeb.trieste.it/domain. Mirror site is available at

http://sbase.abc.hu/domain. A standalone programme will be available on request.

SUPPLEMENTARY INFORMATION: A series of help files is available at the above

addresses.

PMID: 10320404 [Indexed for MEDLINE]

3752. Bioinformatics. 1999 Apr;15(4):341-2.

MIAH: automatic alignment of eukaryotic SSU rRNAs.

Thébault P(1), Monestié P, McGrath A, Higgins DG.

Author information:

(1)Department of Biochemistry, University College, Cork, Ireland.

SUMMARY: MIAH is a WWW server for the automatic alignment of new eukaryotic SSU

rRNA sequences to an existing alignment of 1500 sequences.

AVAILABILITY: http://chah.ucc.ie/MIAH Contact :

PMID: 10320403 [Indexed for MEDLINE]

3753. Biophys J. 1999 Apr;76(4):2230-7.

Toward objective selection of representative microscope images.

Markey MK(1), Boland MV, Murphy RF.

Author information:

(1)Center for Light Microscope Imaging and Biotechnology, Carnegie Mellon

University, Pittsburgh, Pennsylvania 15213, USA.

Scientists wishing to communicate the essential characteristics of a pattern

(such as an immunofluorescence distribution) currently must make a subjective

choice of one or two images to publish. We therefore developed methods for

objectively choosing a typical image from a set, with emphasis on images from

cell biology. The methods involve calculation of numerical features to describe

each image, calculation of similarity between images as a distance in feature

space, and ranking of images by distance from the center of the feature

distribution. Two types of features were explored, image texture measures and

Zernike polynomial moments, and various distance measures were utilized. Criteria

for evaluating methods for assigning typicality were proposed and applied to sets

of images containing more than one pattern. The results indicate the importance

of using distance measures that are insensitive to the presence of outliers. For

collections of images of the distributions of a lysosomal protein, a Golgi

protein, and nuclear DNA, the images chosen as most typical were in good

agreement with the conventional understanding of organelle morphologies. The

methods described here have been implemented in a web server

(http://murphylab.web.cmu.edu/services/TyplC).

DOI: 10.1016/S0006-3495(99)77379-0

PMCID: PMC1300196

PMID: 10096918 [Indexed for MEDLINE]

3754. Electrophoresis. 1999 Apr-May;20(4-5):891-7.

A two-dimensional electrophoresis database of rat heart proteins.

Li XP(1), Pleissner KP, Scheler C, Regitz-Zagrosek V, Salnikow J, Jungblut PR.

Author information:

(1)Max Volmer Institute of Biophysical Chemistry and Biochemistry, Technical

University Berlin, Germany.

More than 3000 myocardial protein species of Wistar Kyoto rat, an important

animal model, were separated by high-resolution two-dimensional gel

electrophoresis (2-DE) and characterized in terms of isoelectric point (pI) and

molecular mass (Mr). Currently, the 2-DE database contains 64 identified

proteins; forty-three were identified by peptide mass fingerprinting (PMF) using

matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS), nine by

exclusive comparison with other 2-DE heart protein databases, and in only 12

cases of 60 attempts N-terminal sequencing was successful. We used the Make2ddb

software package downloaded from the ExPASy server for the construction of a rat

myocardial 2-DE database. The Make2ddb package simplifies the creation of a new

2-DE database if the Melanie II software and a Sun workstation under Solaris are

available. Our 2-DE database of rat heart proteins can be accessed at URL

http://gelmatching.inf.fu-berlin.de/pleiss/2d.

DOI: 10.1002/(SICI)1522-2683(19990101)20:4/5<891::AID-ELPS891>3.0.CO;2-2

PMID: 10344264 [Indexed for MEDLINE]

3755. Electrophoresis. 1999 Apr-May;20(4-5):755-65.

New algorithmic approaches to protein spot detection and pattern matching in

two-dimensional electrophoresis gel databases.

Pleissner KP(1), Hoffmann F, Kriegel K, Wenk C, Wegner S, Sahlström A, Oswald H,

Alt H, Fleck E.

Author information:

(1)Department of Internal Medicine/Cardiology, Charité, Campus Virchow-Clinic,

Humboldt University and German Heart Institute, Berlin. pleiss@dhzb.de

Protein spot identification in two-dimensional electrophoresis gels can be

supported by the comparison of gel images accessible in different World Wide Web

two-dimensional electrophoresis (2-DE) gel protein databases. The comparison may

be performed either by visual cross-matching between gel images or by automatic

recognition of similar protein spot patterns. A prerequisite for the automatic

point pattern matching approach is the detection of protein spots yielding the

x(s),y(s) coordinates and integrated spot intensities i(s). For this purpose an

algorithm is developed based on a combination of hierarchical watershed

transformation and feature extraction methods. This approach reduces the strong

over-segmentation of spot regions normally produced by watershed transformation.

Measures for the ellipticity and curvature are determined as features of spot

regions. The resulting spot lists containing x(s),y(s),i(s)-triplets are

calculated for a source as well as for a target gel image accessible in 2-DE gel

protein databases. After spot detection a matching procedure is applied. Both the

matching of a local pattern vs. a full 2-DE gel image and the global matching

between full images are discussed. Preset slope and length tolerances of pattern

edges serve as matching criteria. The local matching algorithm relies on a data

structure derived from the incremental Delaunay triangulation of a point set and

a two-step hashing technique. For the incremental construction of triangles the

spot intensities are considered in decreasing order. The algorithm needs neither

landmarks nor an a priori image alignment. A graphical user interface for spot

detection and gel matching is written in the Java programming language for the

Internet. The software package called CAROL (http://gelmatching.inf.fu-berlin.de)

is realized in a client-server architecture.

DOI: 10.1002/(SICI)1522-2683(19990101)20:4/5<755::AID-ELPS755>3.0.CO;2-6

PMID: 10344245 [Indexed for MEDLINE]

3756. J Struct Biol. 1999 Apr-May;125(2-3):112-22.

A framework for querying a database for structural information on 3D images of

macromolecules: A web-based query-by-content prototype on the BioImage

macromolecular server.

de Alarcón PA(1), Gupta A, Carazo JM.

Author information:

(1)Centro Nacional de Biotecnología-CSIC, Campus Universidad Autonoma,

Cantoblanco, Madrid, 28049, Spain.

Nowadays we are experiencing a remarkable growth in the number of databases that

have become accessible over the Web. However, in a certain number of cases, for

example, in the case of BioImage, this information is not of a textual nature,

thus posing new challenges in the design of tools to handle these data. In this

work, we concentrate on the development of new mechanisms aimed at "querying"

these databases of complex data sets by their intrinsic content, rather than by

their textual annotations only. We concentrate our efforts on a subset of

BioImage containing 3D images (volumes) of biological macromolecules,

implementing a first prototype of a "query-by-content" system. In the context of

databases of complex data types the term query-by-content makes reference to

those data modeling techniques in which user-defined functions aim at

"understanding" (to some extent) the informational content of the data sets. In

these systems the matching criteria introduced by the user are related to

intrinsic features concerning the 3D images themselves, hence, complementing

traditional queries by textual key words only. Efficient computational algorithms

are required in order to "extract" structural information of the 3D images prior

to storing them in the database. Also, easy-to-use interfaces should be

implemented in order to obtain feedback from the expert. Our query-by-content

prototype is used to construct a concrete query, making use of basic structural

features, which are then evaluated over a set of three-dimensional images of

biological macromolecules. This experimental implementation can be accessed via

the Web at the BioImage server in Madrid, at

http://www.bioimage.org/qbc/index.html.

Copyright 1999 Academic Press.

DOI: 10.1006/jsbi.1999.4102

PMID: 10222268 [Indexed for MEDLINE]

3757. J Struct Biol. 1999 Apr-May;125(2-3):103-11.

Design and realization of an on-line database for multidimensional microscopic

images of biological specimens.

Lindek S(1), Fritsch R, Machtynger J, de Alarcón PA, Chagoyen M.

Author information:

(1)European Molecular Biology Laboratory (EMBL), Heidelberg, D-69012, Germany.

The BioImage database is a new scientific database for multidimensional

microscopic images of biological specimens, which is available through the World

Wide Web (WWW). The development of this database has followed an iterative

approach, in which requirements and functionality have been revised and extended.

The complexity and innovative use of the data meant that technical and biological

expertise has been crucial in the initial design of the data model. A controlled

vocabulary was introduced to ensure data consistency. Pointers are used to

reference information stored in other databases. The data model was built using

InfoModeler as a database design tool. The database management system is the

Informix Dynamic Server with Universal Data Option. This object-relational system

allows the handling of complex data using features such as collection types,

inheritance, and user-defined data types. Informix datablades are used to provide

additional functionality: the Web Integration Option enables WWW access to the

database; the Video Foundation Blade provides functionality for video handling.

Copyright 1999 Academic Press.

DOI: 10.1006/jsbi.1999.4092

PMID: 10222267 [Indexed for MEDLINE]

3758. Bioinformatics. 1999 Mar;15(3):262-3.

WWW access to the SYSTERS protein sequence cluster set.

Krause A(1), Nicodème P, Bornberg-Bauer E, Rehmsmeier M, Vingron M.

Author information:

(1)Deutsches Krebsforschungszentrum, Theoretische Bioinformatik, INF 280, D-69120

Heidelberg, Germany. A.Krause@DKFZ-Heidelberg.de

SUMMARY: We present a Web server where the SYSTERS cluster set of the

non-redundant protein database consisting of sequences from SWISS-PROT and PIR is

being made available for querying and browsing. The cluster set can be searched

with a new sequence using the SSMAL search tool. Additionally, a multiple

alignment is generated for each cluster and annotated with domain information

from the Pfam protein family database.

AVAILABILITY: The server address is

http://www.dkfz-heidelberg.de/tbi/services/cluster/ systersform

PMID: 10222416 [Indexed for MEDLINE]

3759. Bioinformatics. 1999 Mar;15(3):211-8.

DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence

alignment.

Morgenstern B(1).

Author information:

(1)GSF - National Research Center for Environment and Health, Institute of

Biomathematics and Biometry, Ingolstädter Landstrasse 1, 85764 Neuherberg,

Germany. burkhard.morgenstern@rp-rorer.co.uk

MOTIVATION: The performance and time complexity of an improved version of the

segment-to-segment approach to multiple sequence alignment is discussed. In this

approach, alignments are composed from gap-free segment pairs, and the score of

an alignment is defined as the sum of so-called weights of these segment pairs.

RESULTS: A modification of the weight function used in the original version of

the alignment program DIALIGN has two important advantages: it can be applied to

both globally and locally related sequence sets, and the running time of the

program is considerably improved. The time complexity of the algorithm is

discussed theoretically, and the program running time is reported for various

test examples.

AVAILABILITY: The program is available on-line at the Bielefeld University

Bioinformatics Server (BiBiServ) http://bibiserv.TechFak.Uni-Bielefeld.DE/dial

ign/

PMID: 10222408 [Indexed for MEDLINE]

3760. Genome Res. 1999 Mar;9(3):277-81.

DNA sequence chromatogram browsing using JAVA and CORBA.

Parsons JD(1), Buehler E, Hillier L.

Author information:

(1)EMBL-Outstation-The European Bioinformatics Institute (EBI), Wellcome Trust

Genome Campus, Hinxton, Cambridge CB10 1SD, UK. jparsons@ebi.ac.uk

DNA sequence chromatograms (traces) are the primary data source for all

large-scale genomic and expressed sequence tags (ESTs) sequencing projects.

Access to the sequencing trace assists many later analyses, for example contig

assembly and polymorphism detection, but obtaining and using traces is

problematic. Traces are not collected and published centrally, they are much

larger than the base calls derived from them, and viewing them requires the

interactivity of a local graphical client with local data. To provide efficient

global access to DNA traces, we developed a client/server system based on

flexible Java components integrated into other applications including an applet

for use in a WWW browser and a stand-alone trace viewer. Client/server

interaction is facilitated by CORBA middleware which provides a well-defined

interface, a naming service, and location independence. [The software is packaged

as a Jar file available from the following URL: http://www.ebi.ac.uk/jparsons.

Links to working examples of the trace viewers can be found at

http://corba.ebi.ac.uk/EST. All the Washington University mouse EST traces are

available for browsing at the same URL.]

PMCID: PMC310717

PMID: 10077534 [Indexed for MEDLINE]

3761. Orthopade. 1999 Mar;28(3):277-284. doi: 10.1007/PL00003608.

The importance of computer-based procedures for planning and documentation of

orthopaedic surgery.

Basad E(1).

Author information:

(1)Orthopädische Klinik des Klinikums, Justus-Liebig-Universität Gießen, Germany.

The demand for efficiency in OR management and increase in the necessity of

surgical documentation require the use of software applications in hospitals. A

client-server based OP-planning and documentation system has been in use in the

department of orthopedic surgery in Giessen University since 1992 and is being

continously further developed. Aside from the lawful requirements, the demands of

clinical doctors have been especially considered. The main functions are

management of non medical patient data, scheduling and documentation of

operations with coding of diagnoses and therapy, tissue banking, implant

inventory, on call scheduling, storage of medical video images, clinical word

processing and e-mail. With an integrated web-server, MedXS has the capabilities

to offer functions accessible over any webbrowser (Netscape(TM),

Internet-Explorer(TM)) in the internet or intranet. Through the usage of this

application clinical procedures could be more efficiently realized and better

agreeing positions with the insurance companies could be reached.

DOI: 10.1007/PL00003608

PMID: 28246946

3762. Bioinformatics. 1999 Feb;15(2):131-40.

FORESST: fold recognition from secondary structure predictions of proteins.

Di Francesco V(1), Munson PJ, Garnier J.

Author information:

(1)Analytical Biostatistics Section, Mathematical and Statistical Computing

Laboratory, The Institute, Center for Information Technology, National Institutes

of Health, Bethesda, MD 20892-5626, USA.

MOTIVATION: A method for recognizing the three-dimensional fold from the protein

amino acid sequence based on a combination of hidden Markov models (HMMs) and

secondary structure prediction was recently developed for proteins in the

Mainly-Alpha structural class. Here, this methodology is extended to Mainly-Beta

and Alpha-Beta class proteins. Compared to other fold recognition methods based

on HMMs, this approach is novel in that only secondary structure information is

used. Each HMM is trained from known secondary structure sequences of proteins

having a similar fold. Secondary structure prediction is performed for the amino

acid sequence of a query protein. The predicted fold of a query protein is the

fold described by the model fitting the predicted sequence the best.

RESULTS: After model cross-validation, the success rate on 44 test proteins

covering the three structural classes was found to be 59%. On seven fold

predictions performed prior to the publication of experimental structure, the

success rate was 71%. In conclusion, this approach manages to capture important

information about the fold of a protein embedded in the length and arrangement of

the predicted helices, strands and coils along the polypeptide chain. When a more

extensive library of HMMs representing the universe of known structural families

is available (work in progress), the program will allow rapid screening of

genomic databases and sequence annotation when fold similarity is not detectable

from the amino acid sequence.

AVAILABILITY: FORESST web server at http://absalpha.dcrt.nih.gov:8008/ for the

library of HMMs of structural families used in this paper. FORESST web server at

http://www.tigr.org/ for a more extensive library of HMMs (work in progress).

CONTACT: valedf@tigr.org; munson@helix.nih.gov; garnier@helix.nih.gov

PMID: 10089198 [Indexed for MEDLINE]

3763. J Med Syst. 1999 Feb;23(1):73-6.

A keyword search system for medical doctors to introduce their patients to

specialists.

Hanada E(1), Ise K, Antoku Y, Matsumura K, Kenjo Y, Koga M, Kashiwagi S, Nose Y.

Author information:

(1)Department of Medical Information Science, Faculty of Medicine, Kyushu

University, Fukuoka, Japan.

We developed an online keyword search system on a WWW server to help medical

doctors to introduce their patients to specialists. The system stores special

knowledge and advanced techniques of 287 medical specialists in Kyushu University

Hospital. The data set was presented by the medical specialists themselves.

Specialists can rewrite their data set using the browser anytime to keep their

data set up to date. In addition, the specialists are reminded to keep their data

set up to date by direct mail once a year. The system does not use any database

management systems. We used only two text files and the "grep" command, one of

the basic commands of the UNIX system. Then, it takes less than 1 sec to search

one medical specialist's data set from 200 specialists' data sets. The system is

useful not only inside the hospital but also for practitioners in the local area.

PMID: 10321381 [Indexed for MEDLINE]

3764. Gene. 1999 Jan 8;226(1):129-37.

Protein-coding regions prediction combining similarity searches and conservative

evolutionary properties of protein-coding sequences.

Rogozin IB(1), D'Angelo D, Milanesi L.

Author information:

(1)Istituto di Tecnologie Biomediche Avanzate CNR, via Fratelli Cervi 93, 20090

Segrate, Milan, Italy.

The gene identification procedure in a completely new gene with no good homology

with protein sequences can be a very complex task. In order to identify the

protein-coding region, a new method, 'SYNCOD', based on the analysis of

conservative evolutionary properties of coding regions, has been realized. This

program is able to identify and use the coding region homologies of the

non-annotated (unknown) protein-coding sequences already present in the

nucleotide sequence databases by using the alignment produced by BLASTN. The

ratio of number mismatches resulting in synonymous codons to the number of

mismatches resulting in non-synonymous codons is estimated for each open reading

frame. Monte Carlo simulations are then used to estimate the significance of the

ratio deviation from random behavior. The SYNCOD program has been tested on

generated random sequences and on different control sets. The high accuracy of

predicting protein-coding regions (the correlation coefficient, CC, varies from

0.67 to 0.79) and the high specificity (the portion of wrong exons, WE, varies

from 0.06 to 0.07) have proved to be important features of the suggested

approach. The SYNCOD program is resident on the ITBA-CNR Web Server and can be

used via the Internet (URL: www.itba.mi.cnr.it/webgene).

PMID: 9889348 [Indexed for MEDLINE]

3765. Bioinformatics. 1999 Jan;15(1):32-7.

RNA movies: visualizing RNA secondary structure spaces.

Evers D(1), Giegerich R.

Author information:

(1)Technische Fakultät, Universität Bielefeld, Postfach 100 131, 33501 Bielefeld,

Germany.dirk@TechFak.Uni-Bielefeld.DE

MOTIVATION: RNA Movies is a system for the visualization of RNA secondary

structure spaces. Its input is a script consisting of primary and secondary

structure information. From this script, the system fully automatically generates

animated graphical structure representations. In this way, it creates the

impression of an RNA molecule exploring its own two-dimensional structure space.

RESULTS: RNA Movies has been used to generate animations of a switching structure

in the spliced leader RNA of Leptomonas collosoma and sequential foldings of

potato spindle tuber viroid transcripts.

AVAILABILITY: Demonstrations of the animations mentioned in this paper can be

viewed on our Bioinformatics web server under the following address:

http://BiBiServ.TechFak.Uni-Bielefeld. DE/rnamovies/. The RNA Movies software is

available upon request from the authors.

PMID: 10068690 [Indexed for MEDLINE]

3766. Genetica. 1999;106(1-2):63-73.

Sequence-dependent modelling of local DNA bending phenomena: curvature prediction

and vibrational analysis.

Vlahovicek K(1), Munteanu MG, Pongor S.

Author information:

(1)International Center for Genetic Engineering and Biotechnology, Trieste,

Italy.

Bending is a local conformational micropolymorphism of DNA in which the original

B-DNA structure is only distorted but not extensively modified. Bending can be

predicted by simple static geometry models as well as by a recently developed

elastic model that incorporate sequence dependent anisotropic bendability (SDAB).

The SDAB model qualitatively explains phenomena including affinity of protein

binding, kinking, as well as sequence-dependent vibrational properties of DNA.

The vibrational properties of DNA segments can be studied by finite element

analysis of a model subjected to an initial bending moment. The frequency

spectrum is obtained by applying Fourier analysis to the displacement values in

the time domain. This analysis shows that the spectrum of the bending vibrations

quite sensitively depends on the sequence, for example the spectrum of a curved

sequence is characteristically different from the spectrum of straight sequence

motifs of identical basepair composition. Curvature distributions are

genome-specific, and pronounced differences are found between protein-coding and

regulatory regions, respectively, that is, sites of extreme curvature and/or

bendability are less frequent in protein-coding regions. A WWW server is set up

for the prediction of curvature and generation of 3D models from DNA sequences

(http:@www.icgeb.trieste.it/dna).

PMID: 10710711 [Indexed for MEDLINE]

3767. Int J Med Inform. 1999 Jan;53(1):79-90.

Web-based training: a new paradigm in computer-assisted instruction in medicine.

Haag M(1), Maylein L, Leven FJ, Tönshoff B, Haux R.

Author information:

(1)Laboratory, Computer-Assisted Instruction in Medicine, University of

Heidelberg, Germany.

Computer-assisted instruction (CAI) programs based on internet technologies,

especially on the world wide web (WWW), provide new opportunities in medical

education. The aim of this paper is to examine different aspects of such

programs, which we call 'web-based training (WBT) programs', and to differentiate

them from conventional CAI programs. First, we will distinguish five different

interaction types: presentation; browsing; tutorial dialogue; drill and practice;

and simulation. In contrast to conventional CAI, there are four architectural

types of WBT programs: client-based; remote data and knowledge; distributed

teaching; and server-based. We will discuss the implications of the different

architectures for developing WBT software. WBT programs have to meet other

requirements than conventional CAI programs. The most important tools and

programming languages for developing WBT programs will be listed and assigned to

the architecture types. For the future, we expect a trend from conventional CAI

towards WBT programs.

PMID: 10075132 [Indexed for MEDLINE]

3768. Nucleic Acids Res. 1999 Jan 1;27(1):171-3.

A new version of the RDP (Ribosomal Database Project).

Maidak BL(1), Cole JR, Parker CT Jr, Garrity GM, Larsen N, Li B, Lilburn TG,

McCaughey MJ, Olsen GJ, Overbeek R, Pramanik S, Schmidt TM, Tiedje JM, Woese CR.

Collaborators: Woese CR(2).

Author information:

(1)Department of Microbiology, Michigan State University, 294 Giltner Hall, East

Lansing, MI 48824-1101, USA. (2)U IL, Urbana

The Ribosomal Database Project (RDP-II), previously described by Maidak et al. [

Nucleic Acids Res. (1997), 25, 109-111], is now hosted by the Center for

Microbial Ecology at Michigan State University. RDP-II is a curated database that

offers ribosomal RNA (rRNA) nucleotide sequence data in aligned and unaligned

forms, analysis services, and associated computer programs. During the past two

years, data alignments have been updated and now include >9700 small subunit rRNA

sequences. The recent development of an ObjectStore database will provide more

rapid updating of data, better data accuracy and increased user access. RDP-II

includes phylogenetically ordered alignments of rRNA sequences, derived

phylogenetic trees, rRNA secondary structure diagrams, and various software

programs for handling, analyzing and displaying alignments and trees. The data

are available via anonymous ftp (ftp.cme.msu. edu) and WWW

(http://www.cme.msu.edu/RDP). The WWW server provides ribosomal probe checking,

approximate phylogenetic placement of user-submitted sequences, screening for

possible chimeric rRNA sequences, automated alignment, and a suggested placement

of an unknown sequence on an existing phylogenetic tree. Additional utilities

also exist at RDP-II, including distance matrix, T-RFLP, and a Java-based viewer

of the phylogenetic trees that can be used to create subtrees.

PMCID: PMC148126

PMID: 9847171 [Indexed for MEDLINE]

3769. Nucleic Acids Res. 1999 Jan 1;27(1):310-1.

The ENZYME data bank in 1999.

Bairoch A(1).

Author information:

(1)Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel

Servet, 1211 Geneva 4, Switzerland. amos.bairoch@medecine.unige.ch

The ENZYME data bank is a repository of information related to the nomenclature

of enzymes. In recent years it has become an indispensable resource for the

development of metabolic databases. The current version contains information on

3704 enzymes. It is available through the ExPASy WWW server

(http://www.expasy.ch/).

PMCID: PMC148167

PMID: 9847212 [Indexed for MEDLINE]

3770. Nucleic Acids Res. 1999 Jan 1;27(1):289-91.

The SWISS-2DPAGE database: what has changed during the last year.

Hoogland C(1), Sanchez JC, Tonella L, Bairoch A, Hochstrasser DF, Appel RD.

Author information:

(1)Swiss Institute of Bioinformatics, c/o Medical Informatics Division, Geneva

University Hospital, 24 rue Micheli-du-Crest, 1211 Geneva 14, Switzerland.

christine.hoogland@isb-sib.ch

SWISS-2DPAGE (http://www.expasy.ch/ch2d/) is an annotated two-dimensional

polyacrylamide gel electrophoresis (2-D PAGE) database established in 1993. The

current release contains 21 reference maps from human and mouse biological

samples, as well as from Saccharomyces cerevisiae, Escherichia coli and

Dictyostelium discoideum origin. These reference maps now have 2480 identified

spots, corresponding to 528 separate protein entries in the database, in addition

to virtual entries for each SWISS-PROT sequence. During the last year, the

SWISS-2DPAGE has undergone major changes. Six new maps have been added, and new

functions to access the data have been provided through the ExPASy server.

Finally, an important change concerns the database funding source.

PMCID: PMC148159

PMID: 9847204 [Indexed for MEDLINE]

3771. Nucleic Acids Res. 1999 Jan 1;27(1):272-4.

ProClass Protein Family Database.

Wu CH(1), Shivakumar S, Huang H.

Author information:

(1)Department of Epidemiology/Biomathematics, University of Texas Health Center

at Tyler, Tyler, TX 75710, USA. wu@uthct.edu

ProClass is a protein family database that organizes non-redundant sequence

entries into families defined collectively by PROSITE patterns and PIR

superfamilies. By combining global similarities and functional motifs into a

single classification scheme, ProClass helps to reveal domain and family

relationships and classify multi-domain proteins. The database currently consists

of more than 120 000 sequence entries, approximately 60% of which is classified

into about 3500 families. To maximize family information retrieval, the database

provides links to various protein family/domain and structural class databases

and contains multiple motif alignments of all PROSITE patterns as well as global

alignments of PIR superfamilies. The motif sequences are retrieved from both

PIR-International and SWISS-PROT databases, including a large number of new

members detected by our GeneFIND family identification system. ProClass can be

used to support full-scale genomic annotation, because of its high classification

rate. The ProClass database is available for on-line search and record retrieval

from our WWW server at http://diana.uthct.edu/proclass.html

PMCID: PMC148154

PMID: 9847199 [Indexed for MEDLINE]

3772. Nucleic Acids Res. 1999 Jan 1;27(1):263-7.

Recent improvements of the ProDom database of protein domain families.

Corpet F(1), Gouzy J, Kahn D.

Author information:

(1)Laboratoire de Génétique Cellulaire, INRA, BP27, F-31326 Castanet-Tolosan

cedex, France. fcorpet@toulouse.inra.fr

The ProDom database contains protein domain families generated from the

SWISS-PROT database by automated sequence comparisons. The current version was

built with a new improved procedure based on recursive PSI-BLAST homology

searches. ProDom can be searched on the World Wide Web to study domain

arrangements within either known families or new proteins, with the help of a

user-friendly graphical interface (http://www.toulouse.inra.fr/prodom.html).

Recent improvements to the ProDom server include: ProDom queries under the SRS

Sequence Retrieval System; links to the PredictProtein server; phylogenetic trees

and condensed multiple alignments for a better representation of large domain

families, with zooming in and out capabilities. In addition, a similar server was

set up to display the outcome of whole genome domain analysis as applied to 17

completed microbial genomes (http://www.toulouse.inra.fr/prodomCG.html ).

PMCID: PMC148152

PMID: 9847197 [Indexed for MEDLINE]

3773. Nucleic Acids Res. 1999 Jan 1;27(1):257-9.

The SBASE protein domain library, release 6.0: a collection of annotated protein

sequence segments.

Murvai J(1), Vlahovicek K, Barta E, Szepesvári C, Acatrinei C, Pongor S.

Author information:

(1)International Centre for Genetic Engineering and Biotechnology, Area Science

Park, 34012 Trieste, Italy.

The sixth release of the SBASE protein domain library sequences contains 130 703

annotated and crossreferenced entries corresponding to structural, functional,

ligand-binding and topogenic segments of proteins. The entries were grouped based

on standard names (2312 groups) and futher classified on the basis of the BLAST

similarity (2463 clusters). Automated searching with BLAST and a new

sequence-plot representation of local domain similarities are available at the

WWW-server http://www.icgeb.trieste.it/sbase. A mirror site is at

http://sbase.abc.hu/sbase. The database is freely available by anonymous 'ftp'

file transfer from ftp.icgeb.trieste.it

PMCID: PMC148150

PMID: 9847195 [Indexed for MEDLINE]

3774. Nucleic Acids Res. 1999 Jan 1;27(1):248-50.

INFOGENE: a database of known gene structures and predicted genes and proteins in

sequences of genome sequencing projects.

Solovyev VV(1), Salamov AA.

Author information:

(1)The Sanger Centre, Hinxton, Cambridge CB10 1SA, UK. solovyev@sanger.ac.uk

INFOGENE is a database of known and predicted gene structures with descriptions

of basic functional signals and gene components. It provides a possibility to

create compilations of sequences with a given gene feature as well as to

accumulate and analyze predicted genes in finished and unfinished sequences from

genome sequencing projects. Protein sequence similarity searches in the database

of predicted proteins is offered through the BLASTP program. INFOGENE is realized

under the Sequence Retrieval System that provides useful links with the other

informational databases. The database is available through the WWW server of the

Computational Genomics Group at http://genomic.sanger.ac.uk/db.html

PMCID: PMC148147

PMID: 9847192 [Indexed for MEDLINE]

3775. Nucleic Acids Res. 1999 Jan 1;27(1):220-5.

PRINTS prepares for the new millennium.

Attwood TK(1), Flower DR, Lewis AP, Mabey JE, Morgan SR, Scordis P, Selley JN,

Wright W.

Author information:

(1)School of Biological Sciences, University of Manchester, Manchester M13 9PT,

UK. attwood@biochemistry.ucl.ac.uk

PRINTS is a diagnostic collection of protein fingerprints. Fingerprints exploit

groups of motifs to build characteristic family signatures, offering improved

diagnostic reliability over single-motif approaches by virtue of the mutual

context provided by motif neighbours. Around 1000 fingerprints have now been

created and stored in PRINTS. The September 1998 release (version 20.0), encodes

approximately 5700 motifs, covering a range of globular and membrane proteins,

modular polypeptides and so on. The database is accessible via the DbBrowser Web

Server at http://www.biochem.ucl.ac.uk/bsm/dbbrowser /. In addition to supporting

its continued growth, recent enhancements to the resource include a BLAST server,

and more efficient fingerprint search software, with improved statistics for

estimating the reliability of retrieved matches. Current efforts are focused on

the design of more automated methods for database maintenance; implementation of

an object-relational schema for efficient data management; and integration with

PROSITE, profiles, Pfam and ProDom, as part of the international InterPro

project, which aims to unify protein pattern databases and offer improved tools

for genome analysis.

PMCID: PMC148140

PMID: 9847185 [Indexed for MEDLINE]

3776. Nucleic Acids Res. 1999 Jan 1;27(1):213-4.

Current status of the Asthma and Allergy Database.

Immervoll T(1), Wjst M.

Author information:

(1)GSF National Research Center for Environment and Health, Institute of

Epidemiology, PO Box 1129, D-85758 Neuherberg, Germany.

The database provides an online resource for access to data on the genetics of

asthma and allergy. This report describes the present status of the site.

Currently, a detailed description of 88 linkage studies (7164 linkage positions)

and 72 mutation studies are available. The results can be accessed in table form

or graphically. The database also contains mouse asthma studies and human

homology relationships, gene expression studies and links to relevant patents.

Technical details about the server architecture, database installation, database

construction, database structure and the user interface are explained elsewhere

[Wjst and Immervoll (1998) Bioinformatics, 14, 827-828]. The URL is

http://cooke.gsf.de

PMCID: PMC148138

PMID: 9847183 [Indexed for MEDLINE]

3777. Nucleic Acids Res. 1999 Jan 1;27(1):174-8.

Database on the structure of large subunit ribosomal RNA.

De Rijk P(1), Robbrecht E, de Hoog S, Caers A, Van de Peer Y, De Wachter R.

Author information:

(1)Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1,

B-2610 Antwerpen, Belgium.

The Antwerp database on large subunit ribosomal RNA now contains 607 complete or

nearly complete aligned sequences. The alignment incorporates secondary structure

information for each sequence. Other information about the sequences, such as

literature references, accession numbers and taxonomic information is also

available. Information from the database can be downloaded or searched on the

rRNA WWW Server at URL http://rrna.uia.ac.be/

PMCID: PMC148127

PMID: 9847172 [Indexed for MEDLINE]

3778. Nucleic Acids Res. 1999 Jan 1;27(1):119-22.

The HuGeMap Database: interconnection and visualization of human genome maps.

Barillot E(1), Pook S, Guyon F, Cussat-Blanc C, Viara E, Vaysseix G.

Author information:

(1)GIS Infobiogen, 7 rue Guy Môquet BP 8, 94801 Villejuif cedex, France.

manu@infobiogen.fr

The HuGeMap database stores the major genetic and physical maps of the human

genome. HuGeMap is accessible on the Web at http://www.

infobiogen.fr/services/Hugemap and through a CORBA server. A standard genome map

data format for the interconnection of genome map databases was defined in

collaboration with the EBI. The HuGeMap CORBA server provides this

interconnection using the interface definition language IDL. Two graphical user

interfaces were developed for the visualization of the HuGeMap data: ZoomMap

(http://www.infobiogen.fr/services/zomit/Zoom Map.html) for navigation by zooming

and data transformation via magic lenses, and MappetShow

(http://www.infobiogen.fr/services/Mappet) for visualizing and comparing maps.

PMCID: PMC148110

PMID: 9847155 [Indexed for MEDLINE]

3779. Nucleic Acids Res. 1999 Jan 1;27(1):85-8.

The FlyBase database of the Drosophila Genome Projects and community literature.

FlyBase Consortium(1).

Collaborators: Gelbart WM, Crosby M, Matthews B, Chillemi J, Twombly SR, Emmert

D, Bayraktaroglu L, Smutniak F, Kossida S, Ashburner M, Drysdale RA, Whitfield

EJ, Millburn GH, de Grey A, Kaufman T, Matthews K, Gilbert D, Strelets V,

Grumbling G, Tolstoshev C, Rubin GM, Lewis S, Helt G, Misra S, Harris N,

Brokstein P, Loraine A, Simas D, Benos T.

Author information:

(1)FlyBase, The Biological Laboratories, Harvard University, 16 Divinity Avenue,

Cambridge, MA 02138, USA.

The FlyBase Drosophila genetics database and the public interfaces of the

Berkeley Drosophila Genome Project (BDGP) and European Drosophila Genome Project

(EDGP) are in the process of integrating. At present, the data of these projects

are available from independent, but hyperlinked, WWW sites (FlyBase URL,

http://flybase. bio.indiana.edu/; BDGP URL, http://fruitfly.berkeley.edu/; EDGP

URL, http://edgp.ebi.ac.uk/ ). Because of the considerable overlap of data

classes between the contributions of the Drosophila genome projects and the

Drosophila community, the new and enlarged FlyBase consortium views the

implementation of a single integrated Drosophila genomics/genetics server as

essential to the scientific community. This integration will occur in a stepwise

fashion over the next 1-2 years. In this report, the salient features of the

current databases and how to interrogate and navigate the extensive data sets are

discussed.

PMCID: PMC148103

PMID: 9847148 [Indexed for MEDLINE]

3780. Nucleic Acids Res. 1999 Jan 1;27(1):63-5.

The Enhanced Microbial Genomes Library.

Perrière G(1), Bessières P, Labedan B.

Author information:

(1)Laboratoire de Biométrie, Génétique et Biologie des Populations, Université

Claude Bernard - Lyon 1, 43, boulevard du 11 Novembre 1918, 69622 Villeurbanne

Cedex, France. perriere@biomserv.univ-lyon1.fr

Since the obtention of the complete sequence of Haemophilus influenzae Rd in

1995, the number of bacterial genomes entirely sequenced has regularly increased.

A problem is that the quality of the annotations of these very large sequences is

usually lower than those of the shorter entries encountered in the repository

collections. Moreover, classical sequence database management systems have

difficulties in handling entries of that size. In this context, we have decided

to build the Enhanced Microbial Genomes Library (EMGLib) in which these two

problems are alleviated. This library contains all the complete genomes from

bacteria already sequenced and the yeast genome in GenBank format. The

annotations are improved by the introduction of data on codon usage, gene

orientation on the chromosome and gene families. It is possible to access EMGLib

through two database systems set up on World Wide Web servers: the PBIL server at

http://pbil.univ-lyon1.fr/emglib/emglib. html and the MICADO server at

http://locus.jouy.inra.fr/micado

PMCID: PMC148098

PMID: 9847143 [Indexed for MEDLINE]

3781. Nucleic Acids Res. 1999 Jan 1;27(1):44-8.

MIPS: a database for genomes and protein sequences.

Mewes HW(1), Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D.

Author information:

(1)GSF-Forschungszentrum für Umwelt und Gesundheit, Munich Information Center for

Protein Sequences, am Max-Planck-Instut für Biochemie, Am Klopferspitz 18,

D-82152 Martinsried, Germany. mewes@mips.biochem.mpg.de

The Munich Information Center for Protein Sequences (MIPS-GSF), Martinsried near

Munich, Germany, develops and maintains genome oriented databases. It is

commonplace that the amount of sequence data available increases rapidly, but not

the capacity of qualified manual annotation at the sequence databases. Therefore,

our strategy aims to cope with the data stream by the comprehensive application

of analysis tools to sequences of complete genomes, the systematic classification

of protein sequences and the active support of sequence analysis and functional

genomics projects. This report describes the systematic and up-to-date analysis

of genomes (PEDANT), a comprehensive database of the yeast genome (MYGD), a

database reflecting the progress in sequencing the Arabidopsis thaliana genome

(MATD), the database of assembled, annotated human EST clusters (MEST), and the

collection of protein sequence data within the framework of the PIR-International

Protein Sequence Database (described elsewhere in this volume). MIPS provides

access through its WWW server (http://www.mips.biochem.mpg.de) to a spectrum of

generic databases, including the above mentioned as well as a database of protein

families (PROTFAM), the MITOP database, and the all-against-all FASTA database.

PMCID: PMC148093

PMID: 9847138 [Indexed for MEDLINE]

3782. Nucleic Acids Res. 1999 Jan 1;27(1):39-43.

The PIR-International Protein Sequence Database.

Barker WC(1), Garavelli JS, McGarvey PB, Marzec CR, Orcutt BC, Srinivasarao GY,

Yeh LS, Ledley RS, Mewes HW, Pfeiffer F, Tsugita A, Wu C.

Author information:

(1)Protein Information Resource, National Biomedical Research Foundation,

Washington, DC 20007, USA, USA.

The Protein Information Resource (PIR; http://www-nbrf.georgetown. edu/pir/)

supports research on molecular evolution, functional genomics, and computational

biology by maintaining a comprehensive, non-redundant, well-organized and freely

available protein sequence database. Since 1988 the database has been maintained

collaboratively by PIR-International, an international association of data

collection centers cooperating to develop this resource during a period of

explosive growth in new sequence data and new computer technologies. The PIR

Protein Sequence Database entries are classified into superfamilies, families and

homology domains, for which sequence alignments are available. Full-scale family

classification supports comparative genomics research, aids sequence annotation,

assists database organization and improves database integrity. The PIR WWW server

supports direct on-line sequence similarity searches, information retrieval, and

knowledge discovery by providing the Protein Sequence Database and other

supplementary databases. Sequence entries are extensively cross-referenced and

hypertext-linked to major nucleic acid, literature, genome, structure, sequence

alignment and family databases. The weekly release of the Protein Sequence

Database can be accessed through the PIR Web site. The quarterly release of the

database is freely available from our anonymous FTP server and is also available

on CD-ROM with the accompanying ATLAS database search program.

PMCID: PMC148092

PMID: 9847137 [Indexed for MEDLINE]

3783. Nucleic Acids Res. 1999 Jan 1;27(1):12-7.

GenBank.

Benson DA(1), Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, Wheeler DL.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health,Building 38A, 8600 Rockville Pike, Bethesda, MD

20894, USA. dab@ncbi.nlm.nih.gov

The GenBank (Registered Trademark symbol) sequence database incorporates DNA

sequences from all available public sources, primarily through the direct

submission of sequence data from individual laboratories and from large-scale

sequencing projects. Most submitters use the BankIt (Web) or Sequin programs to

format and send sequence data. Data exchange with the EMBL Data Library and the

DNA Data Bank of Japan helps ensure comprehensive worldwide coverage. GenBank

data is accessible through NCBI's integrated retrieval system, Entrez, which

integrates data from the major DNA and protein sequence databases along with

taxonomy, genome and protein structure information. MEDLINE (Registered Trademark

symbol) s from published articles describing the sequences are included as an

additional source of biological annotation through the PubMed search system.

Sequence similarity searching is offered through the BLAST series of database

search programs. In addition to FTP, Email, and server/client versions of Entrez

and BLAST, NCBI offers a wide range of World Wide Web retrieval and analysis

services based on GenBank data. The GenBank database and related resources are

freely accessible via the URL: http://www.ncbi.nlm.nih.gov

PMCID: PMC148087

PMID: 9847132 [Indexed for MEDLINE]

3784. Nucleic Acids Res. 1999 Jan 1;27(1):10-1.

DBcat: a catalog of biological databases.

Discala C(1), Ninnin M, Achard F, Barillot E, Vaysseix G.

Author information:

(1)GIS Infobiogen, 7 rue Guy Môquet BP 8, 94801 Villejuif cedex, France.

The DBcat (http://www.infobiogen.fr/services/dbcat) is a comprehensive catalog of

biological databases, maintained and curated on a daily basis at GIS Infobiogen.

It contains more than 400 databases classified by application domains. The DBcat

is a structured flat file library, that can be searched by means of an SRS server

or a dedicated Web interface. The files are available for downloading from

Infobiogen anonymous ftp server.

PMCID: PMC148086

PMID: 9847131 [Indexed for MEDLINE]

3785. Proc AMIA Symp. 1999:795-9.

Using external data sources to improve audit trail analysis.

Herting RL Jr(1), Asaro PV, Roth AC, Barnes MR.

Author information:

(1)Department of Health Management and Informatics, University of

Missouri-Columbia School of Medicine, USA.

Audit trail analysis is the primary means of detection of inappropriate use of

the medical record. While audit logs contain large amounts of information, the

information required to determine useful user-patient relationships is often not

present. Adequate information isn't present because most audit trail analysis

systems rely on the limited information available within the medical record

system. We report a feature of the STAR (System for Text Archive and Retrieval)

audit analysis system where information available in the medical record is

augmented with external information sources such as: database sources,

Light-weight Directory Access Protocol (LDAP) server sources, and World Wide Web

(WWW) database sources. We discuss several issues that arise when combining the

information from each of these disparate information sources. Furthermore, we

explain how the enhanced person specific information obtained can be used to

determine user-patient relationships that might signify a motive for

inappropriately accessing a patient's medical record.

PMCID: PMC2232542

PMID: 10566469 [Indexed for MEDLINE]

3786. Proc AMIA Symp. 1999:765-9.

Multimedia explorer: image database, image proxy-server and search-engine.

Frankewitsch T(1), Prokosch U.

Author information:

(1)Department of Medical Informatics and Biomathematics, University of Muenster,

Germany.

Multimedia plays a major role in medicine. Databases containing images, movies or

other types of multimedia objects are increasing in number, especially on the

WWW. However, no good retrieval mechanism or search engine currently exists to

efficiently track down such multimedia sources in the vast of information

provided by the WWW. Secondly, the tools for searching databases are usually not

adapted to the properties of images. HTML pages do not allow complex searches.

Therefore establishing a more comfortable retrieval involves the use of a higher

programming level like JAVA. With this platform independent language it is

possible to create extensions to commonly used web browsers. These applets offer

a graphical user interface for high level navigation. We implemented a database

using JAVA objects as the primary storage container which are then stored by a

JAVA controlled ORACLE8 database. Navigation depends on a structured vocabulary

enhanced by a semantic network. With this approach multimedia objects can be

encapsulated within a logical module for quick data retrieval.

PMCID: PMC2232870

PMID: 10566463 [Indexed for MEDLINE]

3787. Proc Int Conf Intell Syst Mol Biol. 1999:95-105.

Using sequence motifs for enhanced neural network prediction of protein distance

constraints.

Gorodkin J(1), Lund O, Andersen CA, Brunak S.

Author information:

(1)Department of Biotechnology, Technical University of Denmark, Lyngby, Denmark.

gorodkin@cbs.dtu.dk

Correlations between sequence separation (in residues) and distance (in Angstrom)

of any pair of amino acids in polypeptide chains are investigated. For each

sequence separation we define a distance threshold. For pairs of amino acids

where the distance between C alpha atoms is smaller than the threshold, a

characteristic sequence (logo) motif, is found. The motifs change as the sequence

separation increases: for small separations they consist of one peak located in

between the two residues, then additional peaks at these residues appear, and

finally the center peak smears out for very large separations. We also find

correlations between the residues in the center of the motif. This and other

statistical analysis are used to design neural networks with enhanced performance

compared to earlier work. Importantly, the statistical analysis explains why

neural networks perform better than simple statistical data-driven approaches

such as pair probability density functions. The statistical results also explain

characteristics of the network performance for increasing sequence separation.

The improvement of the new network design is significant in the sequence

separation range 10-30 residues. Finally, we find that the performance curve for

increasing sequence separation is directly correlated to the corresponding

information content. A WWW server, distanceP, is available at

http://www.cbs.dtu.dk/services/distanceP/.

PMID: 10786291 [Indexed for MEDLINE]

3788. Radiographics. 1999 Jan-Feb;19(1):169-82.

A radiology department intranet: development and applications.

Willing SJ(1), Berland LL.

Author information:

(1)Department of Radiology, University of Alabama, Birmingham 35233, USA.

An intranet is a "private Internet" that uses the protocols of the World Wide Web

to share information resources within a company or with the company's business

partners and clients. The hardware requirements for an intranet begin with a

dedicated Web server permanently connected to the departmental network. The heart

of a Web server is the hypertext transfer protocol (HTTP) service, which receives

a page request from a client's browser and transmits the page back to the client.

Although knowledge of hypertext markup language (HTML) is not essential for

authoring a Web page, a working familiarity with HTML is useful, as is knowledge

of programming and database management. Security can be ensured by using scripts

to write information in hidden fields or by means of "cookies." Interfacing

databases and database management systems with the Web server and conforming the

user interface to HTML syntax can be achieved by means of the common gateway

interface (CGI), Active Server Pages (ASP), or other methods. An intranet in a

radiology department could include the following types of content: on-call

schedules, work schedules and a calendar, a personnel directory, resident

resources, memorandums and discussion groups, software for a radiology

information system, and databases.

DOI: 10.1148/radiographics.19.1.g99ja20169

PMID: 9925398 [Indexed for MEDLINE]

3789. Rom J Morphol Embryol. 1999-2004;45:3-9.

From telepathology to virtual pathology institution: the new world of digital

pathology.

Kayser K(1), Kayser G, Radziszowski D, Oehmann A.

Author information:

(1)UICC-TPCC, Institute of Pathology, Charité, Berlin, Germany.

klaus.kayser@charite.de

Telepathology has left its childhood. Its technical development is mature, and

its use for primary (frozen section) and secondary (expert consultation)

diagnosis has been expanded to a great amount. This is in contrast to a virtual

pathology laboratory, which is still under technical constraints. Similar to

telepathology, which can also be used for e-learning and e-training in pathology,

as exemplarily is demonstrated on Digital Lung Pathology

(Klaus.Kayser@charite.de) at least two kinds of virtual pathology laboratories

will be implemented in the near future: a) those with distributed pathologists

and distributed (> or = 1) laboratories associated to individual biopsy

stations/surgical theatres, and b) distributed pathologists (usually situated in

one institution) and a centralized laboratory, which digitizes complete

histological slides. Both scenarios are under intensive technical investigations.

The features of virtual pathology comprise a virtual pathology institution (mode

a) that accepts a complete case with the patient's history, clinical findings,

and (pre-selected) images for first diagnosis. The diagnostic responsibility is

that of a conventional institution. The Internet serves as platform for

information transfer, and an open server such as the iPATH

(http://telepath.patho.unibas.ch) for coordination and performance of the

diagnostic procedure. The size and number of transferred images have to be

limited, and usual different magnifications have to be used. The sender needs to

possess experiences in image sampling techniques, which present with the most

significant information. A group of pathologists is "on duty", or selects one

member for a predefined duty period. The diagnostic statement of the

pathologist(s) on duty is retransmitted to the sender with full responsibility.

The first experiences of a virtual pathology institution group working with the

iPATH server working with a small hospital of the Salomon islands are promising.

A centralized virtual pathology institution (mode b) depends upon the

digitalization of a complete slide, and the transfer of large sized images to

different pathologists working in one institution. The technical performance of

complete slide digitalization is still under development. Virtual pathology can

be combined with e-learning and e-training, that will serve for a powerful

daily-work-integrated pathology system. At present, e-learning systems are

"stand-alone" solutions distributed on CD or via Internet. A characteristic

example is the Digital Lung Pathology CD, which includes about 60 different rare

and common lung diseases with some features of electronic communication. These

features include access to scientific library systems (PubMed), distant

measurement servers (EuroQuant), automated immunohisto-chemistry measurements, or

electronic journals (Elec J Pathol Histol, www.pathology-online.org). It combines

e-learning and e-training with some acoustic support. A new and complete database

based upon this CD will combine e-learning and e-teaching with the actual

workflow in a virtual pathology institution (mode a). The technological problems

are solved and do not depend upon technical constraints such as slide scanning

systems. At present, telepathology serves as promoter for a complete new

landscape in diagnostic pathology, the so-called virtual pathology institution.

Industrial and scientific efforts will probably allow an implementation of this

technique within the next two years with exciting diagnostic and scientific

perspectives.

PMID: 15847374 [Indexed for MEDLINE]

3790. Stud Health Technol Inform. 1999;64:217-29.

WWW-based access to radiological patient data: two years of experience.

Bergh B(1), Felix R.

Author information:

(1)Dept. of Information and Data Processing, Medical Centre of the Johann

Wolfgang Goethe-University, Frankfurt am Main, Germany.

This paper describes the theoretical concept behind, the technical implementation

of, and the experiences created with a self-developed system to access

radiological patient data by WWW-technology.MATERIALS AND METHODS: The WWW-system

interfaces the RIS via HL7 and the PACS via DICOM. Core components are a secure

Web-Server and an underlying database system (DB). The DB contains all relevant

information from the RIS and thumbnails of corresponding images archived in the

PACS. After authorization, the user receives this information in the form of

dynamically generated Web pages. By mouse-clicking a thumbnail, the original

DICOM-image is sent. It is displayed with a helper application (DICOM-Viewer),

allowing all kinds of image manipulation and post-processing.

RESULTS: Determined advantages of the system were: platform independence,

security features, a fixed link of image and report, universal availability, and

simple usage. The only critical issue was performance. Cost savings could be seen

in: a reduction of DICOM workstations, employment of all available hardware, and

reduced training and teaching efforts.

CONCLUSIONS: The author believes that a WWW-based concept is the only feasible

approach which is in the same way technically possible, clinically acceptable,

and financially affordable in order to grant a variety of users access to

radiological data--although improvements in performance have to made by, e.g.,

closer implementation of those systems to the RIS and PACS architecture, and by

employing high-grade image compression.

PMID: 10747541 [Indexed for MEDLINE]

3791. Stud Health Technol Inform. 1999;68:568-72.

Generalisation and extension of a web-based data collection system for clinical

studies using Java and CORBA.

Eich HP(1), Ohmann C.

Author information:

(1)Dept. of General and Trauma Surgery, Heinrich-Heine-University Düsseldorf,

Germany. eich@uni-duesseldorf.de

Inadequate informatical support of multi-centre clinical trials lead to pure

quality. In order to support a multi-centre clinical trial a data collection via

WWW and Internet based on Java has been developed. In this study a generalization

and extension of this prototype has been performed. The prototype has been

applied to another clinical trial and a knowledge server based on C+t has been

integrated via CORBA. The investigation and implementation of security aspects of

web-based data collection is now under evaluation.

PMID: 10724953 [Indexed for MEDLINE]

3792. Stud Health Technol Inform. 1999;62:228-34.

BRAVO/TeleTrend: a comprehensive WWW-based neuromonitoring system for the

neurosurgery ICU.

Nenov VI(1), Buxey F, Yamaguchi Y.

Author information:

(1)Brain Monitoring and Modeling Laboratory, UCLA School of Medicine 90095, USA.

This paper describes BRAVO/TeleTrend--a comprehensive client/server-based system

for remote access, review and analyses of continuously acquired multiparametric

physiological data from Intensive Care unit (ICU) patients. The system is

designed as a distributed three tier model and implemented in Java (Sun

Microsystems). TeleTrend is a data review package, which interfaces to existing

physiological bedside monitors such as the BRAVO suite of products (Nicolet

Biomedical, Madison, WI) and the vital signs monitors compatible with the Unity

Network (Marquette Electronics, Milwaukee, WI). It does not transfer over the web

the entire patient record, which can be hundreds of megabytes. Instead, it

provides tools to view a compressed representation of the raw data in a trend

display and to zoom into the raw data if needed. Thus, it eliminates the need for

a high-bandwidth Internet connection and makes possible the use of a slower modem

access to the vast amount of physiological data acquired per patient. In

addition, TeleTrend features a rule-based module capable of generating clinical

alerts, which is a potentially useful tool for neurointensivists and other

critical care personnel. Finally, TeleTrend is intended as a multi-user, semi

real-time telemedical application, which features built-in white-board and chat

components. These components allow several physicians at different locations

around the world to simultaneously view and brainstorm over critical chunks of

continuously recorded raw and trend data. By allowing the end-user user to switch

on-the-fly from monitoring patients in one ICU to those in another, and by

integrating an HL7 interface TeleTrend steps over the boundaries of a single ICU.

Thus, it can be provide a medical enterprise-wide solution to the remote access

of an important component of the electronic patient medical record. Currently in

house validation, verification and alpha testing of the system are underway.

PMID: 10538362 [Indexed for MEDLINE]

3793. Acta Crystallogr D Biol Crystallogr. 1998 Nov 1;54(Pt 6 Pt 1):1155-67.

Classifying a protein in the CATH database of domain structures.

Orengo CA(1), Martin AM, Hutchinson G, Jones S, Jones DT, Michie AD, Swindells

MB, Thornton JM.

Author information:

(1)Biomolecular Structure and Modelling Unit, Department of Biochemistry and

Molecular Biology, Gower St, London WC1E 6BT, England.

orengo@biochemistry.ucl.ac.uk

The CATH database of protein domain structures classifies structures according to

their (C)lass, (A)rchitecture, (T)opology or fold and (H)omologous family

(http://www.biochem.ucl.ac.uk/bsm/cath). Although the protocol used is mostly

automatic, manual inspection is used to check assignments at some critical

stages, such as the detection of very distantly related homologues and anologues

and the assignment of novel architectures. Described in this article is a

recently established facility to search the database with the coordinates of a

newly determined structure. The CATH server first locates domain boundaries and

then uses automatic sequence and structure comparison methods to assign this new

structure to one or more of the domain families within CATH. Diagnostic reports

are generated, together with multiple structural alignments for close relatives.

The Server can be accessed over the World Wide Web (WWW) and mirror sites are

planned to improve access.

PMID: 10089492 [Indexed for MEDLINE]

3794. Artif Intell Med. 1998 Nov;14(3):279-93.

A development environment for knowledge-based medical applications on the

World-Wide Web.

Riva A(1), Bellazzi R, Lanzola G, Stefanelli M.

Author information:

(1)Laboratorio di Informatica Medica, IRCCS Policlinico S. Matteo, Pavia, Italy.

The World-Wide Web (WWW) is increasingly being used as a platform to develop

distributed applications, particularly in contexts, such as medical ones, where

high usability and availability are required. In this paper we propose a

methodology for the development of knowledge-based medical applications on the

web, based on the use of an explicit domain ontology to automatically generate

parts of the system. We describe a development environment, centred on the

LISPWEB Common Lisp HTTP server, that supports this methodology, and we show how

it facilitates the creation of complex web-based applications, by overcoming the

limitations that normally affect the adequacy of the web for this purpose.

Finally, we present an outline of a system for the management of diabetic

patients built using the LISPWEB environment.

PMID: 9821518 [Indexed for MEDLINE]

3795. Genomics. 1998 Nov 1;53(3):325-37.

A database of experimental results on globin gene expression.

Riemer C(1), ElSherbini A, Stojanovic N, Schwartz S, Kwitkin PB, Miller W,

Hardison R.

Author information:

(1)The Center for Gene Regulation, The Pennsylvania State University, University

Park, Pennsylvania, 16802, USA.

Information on gene expression and regulation is expanding rapidly, and the

complexity of the experimental design and data makes unique demands on databases

to store the results. We describe a prototype database containing experimental

results on the expression of mammalian beta-like globin genes, along with several

query methods for accessing the information. The database has tables for DNA

transfer experiments, protein-DNA binding results, and positions of DNase

hypersensitive sites, which make extensive use of nested data structures.

Comparison of data from various mammals is accomplished by providing a common

coordinate system via a simultaneous alignment of matching DNA sequences.

Interactive access to the database is available at a site called the Globin Gene

Server on the World Wide Web (http://globin.cse. psu.edu). This software should

be useful for any genetic system in which DNA sequence data are available.

Copyright 1998 Academic Press.

DOI: 10.1006/geno.1998.5524

PMID: 9799599 [Indexed for MEDLINE]

3796. JAMA. 1998 Oct 21;280(15):1330-2.

FluNet as a tool for global monitoring of influenza on the Web.

Flahault A(1), Dias-Ferrao V, Chaberty P, Esteves K, Valleron AJ, Lavanchy D.

Author information:

(1)Institut National de la Santé et de la Recherche Médicale Unit 444

(Epidémiologie et Sciences de l'Information), Paris, France.

flahault@b3e.jussieu.fr

In collaboration with the Institut National de la Santé et de la Recherche

Médicale, the World Health Organization (WHO) has developed an Internet

application linking the global WHO network of influenza centers (FluNet;

http://oms.b3e.jussieu.fr/flunet/). During 1997, 22 pilot centers entered data on

influenza activity and viral laboratory results directly into FluNet via secured

access. In addition, 54 centers sent data to WHO for entry. Four countries (the

Russian Federation, Romania, Sweden, and the United Kingdom) reported widespread

outbreaks of at least 4 weeks' duration. The FluNet server ran 24 hours a day

without interruption. To improve management and enhance standardization of

reporting, this early-alert system for the global monitoring of influenza

provides international and national authorities, the public, and the media with

full access to real-time epidemiological and virological information.

PMID: 9794312 [Indexed for MEDLINE]

3797. Gene. 1998 Oct 9;221(1):GC57-63.

CINEMA--a novel colour INteractive editor for multiple alignments.

Parry-Smith DJ(1), Payne AW, Michie AD, Attwood TK.

Author information:

(1)Department of Discovery Biology, Pfizer Central Research, Sandwich, UK.

DJParry-Smith@camdd.co.uk

CINEMA is a new editor for manipulating and generating multiple sequence

alignments. The program provides both an interface to existing databases of

alignments on the Internet and a tool for constructing and modifying alignments

locally. It is written in Java, so executable code will run on most major desktop

platforms without modification. The implementation is highly flexible, so the

applet can be easily customised with additional functions; and the object classes

are reusable, promoting rapid development of program extensions. Formerly, such

extended functionality might have been provided via browser plug-ins, which have

to be downloaded and installed on every client before loading data. Now, for the

first time, an applet is available that allows interactive client-side processing

of an alignment, which can then be stored or processed automatically on the

server. The program is embedded in a comprehensive help file and is accessible

both as a stand-alone tool on UCL's Bioinformatics Server;

http:/(/)www.biochem.ucl.ac.uk/bsm/dbbrowser+ ++/CINEMA2.02/, and as an integral

part of the PRINTS protein fingerprint database. Exploitation of such novel

technologies revolutionises the way users may interact with public databases in

the future: bioinformatics centres need not simply provide data, but are now able

to offer the means by which information is visualised and manipulated, without

the requirement for users to install software.

PMID: 9852962 [Indexed for MEDLINE]

3798. Med Inform (Lond). 1998 Oct-Dec;23(4):277-87.

A network-based training environment: a medical image processing paradigm.

Costaridou L(1), Panayiotakis G, Sakellaropoulos P, Cavouras D, Dimopoulos J.

Author information:

(1)Department of Medical Physics, School of Medicine, University of Patras,

Greece.

The capability of interactive multimedia and Internet technologies is

investigated with respect to the implementation of a distance learning

environment. The system is built according to a client-server architecture, based

on the Internet infrastructure, composed of server nodes conceptually modelled as

WWW sites. Sites are implemented by customization of available components. The

environment integrates network-delivered interactive multimedia courses,

network-based tutoring, SIG support, information databases of professional

interest, as well as course and tutoring management. This capability has been

demonstrated by means of an implemented system, validated with digital image

processing content, specifically image enhancement. Image enhancement methods are

theoretically described and applied to mammograms. Emphasis is given to the

interactive presentation of the effects of algorithm parameters on images. The

system end-user access depends on available bandwidth, so high-speed access can

be achieved via LAN or local ISDN connections. Network based training offers new

means of improved access and sharing of learning resources and expertise, as

promising supplements in training.

PMID: 9922949 [Indexed for MEDLINE]

3799. Nucleic Acids Res. 1998 Sep 15;26(18):4280-90.

A database of macromolecular motions.

Gerstein M(1), Krebs W.

Author information:

(1)Department of Molecular Biophysics and Biochemistry, 266 Whitney Avenue, Yale

University, PO Box 208114, New Haven, CT 06520, USA. mark.gerstein@yale.edu

We describe a database of macromolecular motions meant to be of general use to

the structural community. The database, which is accessible on the World Wide Web

with an entry point at http://bioinfo.mbb.yale.edu/MolMovDB , attempts to

systematize all instances of protein and nucleic acid movement for which there is

at least some structural information. At present it contains >120 motions, most

of which are of proteins. Protein motions are further classified hierarchically

into a limited number of categories, first on the basis of size (distinguishing

between fragment, domain and subunit motions) and then on the basis of packing.

Our packing classification divides motions into various categories (shear, hinge,

other) depending on whether or not they involve sliding over a continuously

maintained and tightly packed interface. In addition, the database provides some

indication about the evidence behind each motion (i.e. the type of experimental

information or whether the motion is inferred based on structural similarity) and

attempts to describe many aspects of a motion in terms of a standardized

nomenclature (e.g. the maximum rotation, the residue selection of a fixed core,

etc.). Currently, we use a standard relational design to implement the database.

However, the complexity and heterogeneity of the information kept in the database

makes it an ideal application for an object-relational approach, and we are

moving it in this direction. Specifically, in terms of storing complex

information, the database contains plausible representations for motion pathways,

derived from restrained 3D interpolation between known endpoint conformations.

These pathways can be viewed in a variety of movie formats, and the database is

associated with a server that can automatically generate these movies from

submitted coordinates.

PMCID: PMC147832

PMID: 9722650 [Indexed for MEDLINE]

3800. Comput Biol Med. 1998 Sep;28(5):459-72.

Evolution of web site design: implications for medical education on the Internet.

Chu LF(1), Chan BK.

Author information:

(1)Stanford University School of Medicine, CA 94305, USA.

lchu@leland.stanford.edu

Since its inception, the world wide web (WWW) has possessed the potential for

becoming a 'watershed' medium for conveying complex, structured information

across vast temporal and geographical barriers. In 1995, the MedWorld project

(http:(/)/medworld.stanford.edu) was created at the Stanford University School of

Medicine in an effort to innovate and explore the design process of creating WWW

applications specifically for medical education. Until recently, the evolution of

WWW applications has been mainly driven by technological advances in

client-server technology, enabling or translating traditional modes of

collaborative medical education (e.g. voice, presence, print, motion) into WWW

devices and applications. Many of these applications, while technologically

advanced, lack focused development of interface and interactivity design, which

may enhance learning experiences. WWW applications which incorporate design

innovation in parity with advances in client-server technology have been termed,

'third generation' web sites and have the potential to improve the quality of WWW

applications designed for medical education. This work describes how the MedWorld

project has created a 'third generation' WWW application by utilizing innovation

in information, interface and interactivity design to create innovative WWW

technology for the medical education arena.

PMID: 9861505 [Indexed for MEDLINE]

3801. J Radiol. 1998 Sep;79(9):825-35.

[The virtual university in medicine. Context, concepts, specifications, users'

manual].

[Article in French]

Duvauferrier R(1), Séka LP, Rolland Y, Rambeau M, Le Beux P, Morcet N.

Author information:

(1)Département de Radiologie, Hôpital Sud, Rennes.

regis.duvauferrier@univ-rennes1.fr

The widespread use of Web servers, with the emergence of interactive functions

and the possibility of credit card payment via Internet, together with the

requirement for continuing education and the subsequent need for a computer to

link into the health care network have incited the development of a virtual

university scheme on Internet. The Virtual University of Radiology is not only a

computer-assisted teaching tool with a set of attractive features, but also a

powerful engine allowing the organization, distribution and control of medical

knowledge available in the www.server. The scheme provides patient access to

general information, a secretary's office for enrollment and the Virtual

University itself, with its library, image database, a forum for subspecialties

and clinical case reports, an evaluation module and various guides and help tools

for diagnosis, prescription and indexing. Currently the Virtual University of

Radiology offers diagnostic imaging, but can also be used by other specialties

and for general practice.

PMID: 9791762 [Indexed for MEDLINE]

3802. Methods Inf Med. 1998 Sep;37(3):247-53.

Using a WWW-based mail user agent for secure electronic mail service for health

care users.

Kiuchi T(1), Ohe K, Kaihara S.

Author information:

(1)University Medical Information Network Center, Tokyo, Japan.

kiuchi-tky@umin.u-tokyo.ac.jp

WWW-based user interface is presented for secure electronic mail service for

healthcare users. Using this method, communications between an electronic mail

(WWW) server and users (WWW browsers) can be performed securely using Secure

Socket Layer protocol-based Hypertext Transfer Protocol (SSL-HTTP). The mail can

be encrypted, signed, and sent to the recipients and vice versa on the remote WWW

server. The merit of this method is that many healthcare users can use a secure

electronic mail system easily and immediately, because SSL-compatible WWW

browsers are widely used and this system can be made available simply by

installing a WWW-based mail user agent on a mail server. We implemented a

WWW-based mail user agent which is compatible with PEM-based secure mail and made

it available to about 16,000 healthcare users. We believe this approach is

effective in facilitating secure network-based information exchange among medical

professionals.

PMID: 9787624 [Indexed for MEDLINE]

3803. Electrophoresis. 1998 Aug;19(11):1960-71.

'98 Escherichia coli SWISS-2DPAGE database update.

Tonella L(1), Walsh BJ, Sanchez JC, Ou K, Wilkins MR, Tyler M, Frutiger S, Gooley

AA, Pescaru I, Appel RD, Yan JX, Bairoch A, Hoogland C, Morch FS, Hughes GJ,

Williams KL, Hochstrasser DF.

Author information:

(1)Clinical Chemistry Laboratory, Geneva University Hospital, Switzerland.

luisa.tonella@dim.hcuge.ch

The combination of two-dimensional polyacrylamide gel electrophoresis (2-D PAGE),

computer image analysis and several protein identification techniques allowed the

Escherichia coli SWISS-2DPAGE database to be established. This is part of the

ExPASy molecular biology server accessible through the WWW at the URL address

http://www.expasy.ch/ch2d/ch2d-top.html . Here we report recent progress in the

development of the E. coli SWISS-2DPAGE database. Proteins were separated with

immobilized pH gradients in the first dimension and sodium dodecyl

sulfate-polyacrylamide gel electrophoresis in the second dimension. To increase

the resolution of the separation and thus the number of identified proteins, a

variety of wide and narrow range immobilized pH gradients were used in the first

dimension. Micropreparative gels were electroblotted onto polyvinylidene

difluoride membranes and spots were visualized by amido black staining. Protein

identification techniques such as amino acid composition analysis, gel comparison

and microsequencing were used, as well as a recently described Edman "sequence

tag" approach. Some of the above identification techniques were coupled with

database searching tools. Currently 231 polypeptides are identified on the E.

coli SWISS-2DPAGE map: 64 have been identified by N-terminal microsequencing, 39

by amino acid composition, and 82 by sequence tag. Of 153 proteins putatively

identified by gel comparison, 65 have been confirmed. Many proteins have been

identified using more than one technique. Faster progress in the E. coli proteome

project will now be possible with advances in biochemical methodology and with

the completion of the entire E. coli genome.

DOI: 10.1002/elps.1150191114

PMID: 9740056 [Indexed for MEDLINE]

3804. Med Inform (Lond). 1998 Jul-Sep;23(3):179-85.

A WWW-based information system on resistance of bacteria to antibiotics.

Schindler J(1), Schindler Z, Schindler J Jr.

Author information:

(1)3rd Faculty of Medicine, Charles University, Prague, Czech Republic.

The information system on resistance of bacteria to antibiotics (WARN--World

Antibiotic Resistance Network) is implemented as a WWW server at Charles

University in Prague (http:/(/)www.warn.cas.cz). Its main goal is to give

information about problems of antibiotic resistance of bacteria and to process

data on isolated strains. The WARN web-site contains six main topics. Four of

them form the core of the system: Topics of Interest bring information on

selected timely topics in antibiotic resistance--pneumococci, staphylococci,

beta-lactamases, glycopeptide--and aminoglycoside resistance. Global Monitor

brings references and reports on resistance in the world as well as recommended

method of surveillance. The topic Data contains raw data on strains in particular

countries and hospitals. Data can be viewed in their original form as a list of

records (strains) or processed to provide statistics about the resistance rates

in the selected country or hospital respectively. The topic Search allows one to

search for one or several terms in the whole document. Counts of accessed pages

show, that there is a standing demand for information about the serious problems

of antibiotic therapy of infectious diseases.

PMID: 9785318 [Indexed for MEDLINE]

3805. Med Inform (Lond). 1998 Jul-Sep;23(3):253-64.

A first evaluation of a pedagogical network for medical students at the

University Hospital of Rennes.

Fresnel A(1), Jarno P, Burgun A, Delamarre D, Denier P, Cleret M, Courtin C, Seka

LP, Pouliquen B, Cléran L, Riou C, Leduff F, Lesaux H, Duvauferrier R, Le Beux P.

Author information:

(1)Laboratoire d'Informatique Médicale, Faculté de Médecine, Rennes, France.

A pedagogical network has been developed at University Hospital of Rennes from

1996. The challenge is to give medical information and informatics tools to all

medical students in the clinical wards of the University Hospital. At first, nine

wards were connected to the medical school server which is linked to the

Internet. Client software electronic mail and WWW Netscape on Macintosh

computers. Sever software is set up on Unix SUN providing a local homepage with

selected pedagogical resources. These documents are stored in a DBMS database

ORACLE and queries can be provided by specialty, authors or disease. The students

can access a set of interactive teaching programs or electronic textbooks and can

explore the Internet through the library information system and search engines.

The teachers can send URL and indexation of pedagogical documents and can produce

clinical cases: the database updating will be done by the users. This experience

of using Web tools generated enthusiasm when we first introduced it to students.

The evaluation shows that if the students can use this training early on, they

will adapt the resources of the Internet to their own needs.

PMID: 9785328 [Indexed for MEDLINE]

3806. Radiat Med. 1998 Jul-Aug;16(4):283-7.

Clinical usefulness of the management and delivery of radiation dose-distribution

images using the Internet.

Nakagawa K(1), Onogi Y, Aoki Y, Kozuka T, Ohtomo K.

Author information:

(1)Department of Radiology, Faculty of Medicine, University of Tokyo, Japan.

Dose distribution images in radiation therapy play important roles in the

management of cancer patients. To date, hard copies of these images have been

stored for referral by radiation oncologists as needed. In most cases, these

images are not available to medical personnel outside the radiation oncology

department. We have developed a means to access these dose distribution images

from the hospital via the World-Wide Web (WWW). A screen snapshot of a dose

distribution image on the CRT of a treatment planning unit is copied to the WWW

server and converted to a GIF (graphic interchange format) image. Similarly, we

can register dose volume histograms and digitally reconstructed radiographs (DRR)

on the WWW. Medical personnel can view these images through the WWW browser from

anywhere in the hospital. As a result, radiation oncologists are given detailed

information on target definition in treatment planning by expert physicians. The

system also helps co-medical personnel in understanding dose distribution and

predicting radiation injury. At the same time, it actualizes an electronic

archive of dose distribution images, which is a database for quick and reliable

review, evaluation, and comparison of treatment plans. This technique also

fosters closer relationships among radiation oncologists, physicians, and

co-medical personnel.

PMID: 9814423 [Indexed for MEDLINE]

3807. FEBS Lett. 1998 Jun 23;430(1-2):28-36.

The complete genome of Bacillus subtilis: from sequence annotation to data

management and analysis.

Moszer I(1).

Author information:

(1)Unité de Régulation de l'Expression Génétique, Institut Pasteur, Paris,

France. moszer@pasteur.fr

The completion of the entire 4.2-Mb genome sequence of the gram-positive

bacterium Bacillus subtilis has been a milestone for biological studies on this

model organism. This paper describes bioinformatics work related to this joint

European and Japanese project: methods and strategies for gene annotation and

detection of sequencing errors, using an integrated cooperative computer

environment (Imagene); construction of a specialized database for data management

and a WWW server for data retrieval (SubtiList); DNA sequence analysis, yielding

striking results on oligonucleotide bias, repeated sequences, and codon usage,

all landmarks of evolutionary events shaping the B. subtilis genome.

PMID: 9678589 [Indexed for MEDLINE]

3808. J Theor Biol. 1998 Jun 21;192(4):475-87.

The subclass approach for mutational spectrum analysis: application of the SEM

algorithm.

Glazko GB(1), Milanesi L, Rogozin IB.

Author information:

(1)Institute of Cytology and Genetics, Novosibirsk, Russia.

Analysis and comparison of mutational spectra represents an important problem in

molecular biology. To analyse a mutational spectra we apply an algorithm based on

the SEM subclass approach (Simulation, Expectation, Maximization). The algorithm

tries to classify the mutational sites according to different mutation

probabilities, and each site should belong to one class. Each class is

approximated by binomial distribution and thus any real mutational spectrum is

regarded as a mixture of binomial distributions. The separation process runs

iteratively. Each iteration includes the simulation, maximization and estimation

procedures. To evaluate the quality of the classification results, the X2 test is

used. The algorithm has been checked on random spectra with preset parameters and

on real mutational spectra. As has been shown, 17 out of 19 analysed real

mutational spectra can be divided into two or more classes of sites, of which one

contains hotspots of mutation. For the G:C-->A:T mutational spectra induced by

Sn1 alkylating mutagenes (11 spectra) the classification accuracy was 0.95. To

test different site volumes, each Sn1-induced spectrum was divided into the G-->A

and C-->T spectra. The classification accuracy for these spectra was 0.96. From

the analysis of classification errors it is possible to suggest that at least

part of them cannot be ascribed to the faults of the algorithm but are caused by

some special features of the mutagenesis itself. The results of the real data are

in good relation with existing knowledge. The approach we present is an attempt

to formalize the concept of a "mutational hotspot". The program implementing the

SEM algorithm is available on the Web server

(http:/(/)www.itba.mi.cnr.it/webmutation).

DOI: 10.1006/jtbi.1998.0668

PMID: 9680721 [Indexed for MEDLINE]

3809. Bioinformatics. 1998 Jun;14(5):452-7.

W2H: WWW interface to the GCG sequence analysis package.

Senger M(1), Flores T, Glatting K, Ernst P, Hotz-Wagenblatt A, Suhai S.

Author information:

(1)DKFZ - German Cancer Research Center, Department of Molecular Biophysics, Im

Neuenheimer Feld 280, D-69120 Heidelberg, Germany. m.senger@ebi.ac.uk

MOTIVATION: The user-friendly, graphical X-windows interface (WPI) to the GCG

sequence analysis package can often not be used due to the lack of an X-server on

PC or Macintosh computers. Because Web browsers like Netscape are much more

common on those platforms, we decided to develop W2H, a WWW interface to the GCG

Sequence Analysis Software Package with nearly the same functionality as the

X-windows interface WPI.

RESULTS: The new WWW interface (W2H) to the GCG Sequence Analysis Software

Package (Wisconsin Package) supports modern Web technologies, like client-pull

method, or embedded scripting language, and provides a reasonable platform

independence. The interface is quite comprehensive with advanced features like

sequence selector, search set builder, enzyme chooser, access to sequence

databases, uploading client files to the GCG server or displaying and

manipulating graphical outputs in addition to GCG analysis programs. W2H also

manages secure access to both GCG server and user data. For special environments,

like workshops, conferences and company intranets, there is a special mode

(Intranet mode) with less security constraints. The behaviour of W2H is mostly

controlled by meta-data files describing the applications and giving a base for

dynamic creation of HTML documents. This paper presents mainly the development

approaches used, and architectural design aspects of W2H.

AVAILABILITY: W2H is available by ftp://ftp.ebi.ac. uk/pub/software/unix/w2h or

ftp://genome.dkfz-heidelberg.de/pub/w2h

CONTACT: m.senger@ebi.ac.uk

PMID: 9682058 [Indexed for MEDLINE]

3810. Bioinformatics. 1998 Jun;14(5):423-9.

Removing near-neighbour redundancy from large protein sequence collections.

Holm L(1), Sander C.

Author information:

(1)EMBL-EBI, Cambridge CB10 1SD, UK.

MOTIVATION: To maximize the chances of biological discovery, homology searching

must use an up-to-date collection of sequences. However, the available sequence

databases are growing rapidly and are partially redundant in content. This leads

to increasing strain on CPU resources and decreasing density of first-hand

annotation.

RESULTS: These problems are addressed by clustering closely similar sequences to

yield a covering of sequence space by a representative subset of sequences. No

pair of sequences in the representative set has >90% mutual sequence identity.

The representative set is derived by an exhaustive search for close similarities

in the sequence database in which the need for explicit sequence alignment is

significantly reduced by applying deca- and pentapeptide composition filters. The

algorithm was applied to the union of the Swissprot, Swissnew, Trembl, Tremblnew,

Genbank, PIR, Wormpep and PDB databases. The all-against-all comparison required

to generate a representative set at 90% sequence identity was accomplished in 2

days CPU time, and the removal of fragments and close similarities yielded a size

reduction of 46%, from 260 000 unique sequences to 140 000 representative

sequences. The practical implications are (i) faster homology searches using, for

example, Fasta or Blast, and (ii) unified annotation for all sequences clustered

around a representative. As tens of thousands of sequence searches are performed

daily world-wide, appropriate use of the non-redundant database can lead to major

savings in computer resources, without loss of efficacy.

AVAILABILITY: A regularly updated non-redundant protein sequence database

(nrdb90), a server for homology searches against nrdb90, and a Perl script

(nrdb90.pl) implementing the algorithm are available for academic use from

http://www.embl-ebi.ac. uk/holm/nrdb90.

CONTACT: holm@embl-ebi.ac.uk

PMID: 9682055 [Indexed for MEDLINE]

3811. Bioinformatics. 1998 Jun;14(5):401-6.

Reduced space hidden Markov model training.

Tarnas C(1), Hughey R.

Author information:

(1)Department of Computer Engineering, Jack Baskin School of Engineering,

University of California, Santa Cruz, CA 95064, USA.

MOTIVATION: Complete forward-backward (Baum-Welch) hidden Markov model training

cannot take advantage of the linear space, divide-and-conquer sequence alignment

algorithms because of the examination of all possible paths rather than the

single best path.

RESULTS: This paper discusses the implementation and performance of

checkpoint-based reduced space sequence alignment in the SAM hidden Markov

modeling package. Implementation of the checkpoint algorithm reduced memory usage

from O(mn) to O (m square root n) with only a 10% slowdown for small m and n, and

vast speed-up for the larger values, such as m = n = 2000, that cause excessive

paging on a 96 Mbyte workstation. The results are applicable to other types of

dynamic programming.

AVAILABILITY: A World-Wide Web server, as well as information on obtaining the

Sequence Alignment and Modeling (SAM) software suite, can be found at

http://www.cse.ucsc. edu/research/compbio/sam.html.

CONTACT: rph@cse.ucsc.edu

PMID: 9682053 [Indexed for MEDLINE]

3812. IEEE Trans Inf Technol Biomed. 1998 Jun;2(2):74-9.

Design and development of an interactive medical teleconsultation system over the

World Wide Web.

Bai J(1), Zhang Y, Dai B.

Author information:

(1)Department of Electrical Engineering, School of Life Science and Engineering,

Tsinghua University, Beijing, China.

The objective of the medical teleconsultation system presented in this paper is

to demonstrate the use of the World Wide Web (WWW) for telemedicine and

interactive medical information exchange. The system, which is developed based on

Java, could provide several basic Java tools to fulfill the requirements of

medical applications, including a file manager, data tool, bulletin board, and

digital audio tool. The digital audio tool uses point-to-point structure to

enable two physicians to communicate directly through voice. The others use

multipoint structure. The file manager manages the medical images stored in the

WWW information server, which come from a hospital database. The data tool

supports cooperative operations on the medical data between the participating

physicians. The bulletin board enables the users to discuss special cases by

writing text on the board, send their personal or group diagnostic reports on the

cases, and reorganize the reports and store them in its report file for later

use. The system provides a hardware-independent platform for physicians to

interact with one another as well as to access medical information over the WWW.

PMID: 10719516 [Indexed for MEDLINE]

3813. Int J Med Inform. 1998 Jun;50(1-3):261-5.

The implementation of an integrated on-line health education system at RMIT.

Zylinski J(1), Allan GL, Jamieson P, Maher KP, Green R, Hislop J.

Author information:

(1)Faculty of Biomedical and Health Sciences, RMIT, Melbourne, Victoria,

Australia. jzylinski@rmit.edu.au

The Faculty of Biomedical and Health Sciences at RMIT has been developing an

on-line health education system using a systems thinking approach, to create a

learning environment whose basis is supported by Information Technology (IT). The

centre-piece of this system is the Faculty Learning Centre, which has been

created, both in space and layout, to promote collaborative learning between the

students, so that the educator is physically assimilated with the student body.

This facility is supplemented by the Faculty WWW server, which has been the main

vehicle for course material dissemination to students. To ensure an effective

on-line teaching environment, the position of an on-line facilitator has been

created, whose responsibilities include both the continual evaluation of the

system and the implementation of appropriate system changes. Aspects have

included the production of a staff development training program and extensive

user documentation. This paper discusses the systems thinking approach used to

implement this integrated on-line system, and the establishment of explicit

educational rationales in the use of IT to support learning strategies. Some

examples of the on-line educational programs are also presented.

PMID: 9726521 [Indexed for MEDLINE]

3814. Int J Med Inform. 1998 Jun;50(1-3):123-32.

Developing curriculum in nursing informatics in Europe.

Mantas J(1).

Author information:

(1)University of Athens, Laboratory of Health Informatics, Greece.

jmantas@dn.uoa.gr

The NIGHTINGALE Project (NIGHTINGALE Project: HC1109 DGXIII Contract and

Technical Annex, European Commission, December 1995) which started on the 1st of

January, 1996, after the approval of the European Commission, has a 36 month

duration. It is essential in planning and implementing a strategy in training the

nursing profession in using and applying healthcare information systems.

NIGHTINGALE contributes towards the appropriate use of the developed telematics

infrastructure across Europe by educating and training nurses in a harmonious way

across Europe in the upcoming field of nursing informatics. NIGHTINGALE develops

courseware material based on the curriculum development process using multimedia

technologies. Computer based training software packages in nursing informatics

will be the basis of the training material and the corresponding courses. CD-ROM

based training and reference material will also be provided in the courses

whereas the traditional booklets, teaching material and textbooks can also play

an adequate role in training. NIGHTINGALE will disseminate all information and

courseware material freely to all interested parties through the publications of

the proceedings of the conferences, through the establishment of the world wide

web (WWW) server in nursing informatics for Europe

(http://www.dn.uoa.gr/nightingale), which will become a depository of nursing

information knowledge across Europe as well as a dissemination node of nursing

informatics throughout the European members states for the benefit and welfare of

the European citizen.

PMID: 9726502 [Indexed for MEDLINE]

3815. Toxicology. 1998 May 15;127(1-3):85-95.

US EPA's IRIS pilot program: establishing IRIS as a centralized, peer-reviewed

data base with agency consensus. Integrated Risk Information System.

Mills A(1), Foureman GL.

Author information:

(1)US EPA, National Center for Environmental Assessment, Washington, DC 20460,

USA.

The US EPA's Integrated Risk Information System (IRIS) contains Agency consensus

scientific positions and quantitative values on cancer and noncancer health

effects that may result from lifetime oral or inhalation exposure to specific

chemical substances in the environment. Combined with specific exposure

assessment information, the summary health information in IRIS may be used as a

source in evaluating potential public health risks from environmental

contaminants. IRIS is available to the public via EPA's Internet server at

http://www.epa.gov/iris. Originally developed for internal EPA use, IRIS usage

has broadened since being made publicly available in 1988 to include the private

and public sectors nationally and internationally. Up to 1995, IRIS summaries

were generated from within various EPA Offices and Regions and reviewed by Agency

Workgroups, one for cancer and one for noncancer endpoints, before entry onto

IRIS. In response to the increasing usage and recognition of IRIS and suggestions

for improvement, an IRIS Pilot program was initiated in 1995. The purpose of the

Pilot was 3-fold: To improve efficiency in getting information on to IRIS; to

improve documentation for the positions reported in IRIS summaries, including

applying new methodologies and guidance; and to improve opportunity for public

input including external peer review. A new infrastructure was put in place,

consisting of a cross-Agency team of 'Chemical Managers', a Pilot Program

Manager, and a set of Agency 'Consensus Reviewers'. Cancer and noncancer

assessments were prepared in an integrated fashion for Pilot chemical substances,

documented in 'Toxicological Reviews' and derivative IRIS summaries. Public input

was emphasized via an initial data call and rigorous external peer review. A

final step was Agency-wide consensus review by senior staff scientists

representing EPA's Offices and Regions. EPA's experience with the Pilot is

forming the basis for designing operational aspects of the long-term IRIS

program.

PMID: 9699796 [Indexed for MEDLINE]

3816. Klin Monbl Augenheilkd. 1998 May;212(5):264-7.

[Digital internet-based ophthalmologic image databank].

[Article in German]

Török B(1), Niederberger H, Somorjai Z, Bischoff P.

Author information:

(1)Kantonsspital St. Gallen, Augenklinik.

OBJECTIVE: To manage an ophthalmological image database easily and rationally.

MATERIALS AND METHODS: Fundus, macro and simultaneously recorded Fluorescein and

ICG angiography images were digitized and archived using personal computers and

UNIX workstations. Image data were written to CD-recordables. The CD-s with image

information were stored in a jukebox-server. On demand image information was

converted to dynamic hypertext markup language (html) with a WWW-server.

Information stored on the servers could be observed with browser programs running

on client computers connected to local area network. The communication with

ophthalmologist working outside our hospital was realized by electronic mail.

RESULTS: Different platforms (PC, Mac, workstations, etc.) and operating systems

(Windows 3.x, 95, NT, MacOS, UNIX, etc.) can be used as clients. The

communication with the system is accomplished by standard internet programs

(Internet Explorer, Netscape Navigator, etc.). Thanks to the intuitive graphical

user interface, no special computer knowledge is required to retrieve the stored

data.

CONCLUSION: Our digital image database has many advantages over a conventional

image archive: it is round the clock available, images can be stored and copied

without loss of quality, digital images can be easily integrated in other

applications, it can be used without special computer knowledge, it is expandable

and compatible with all contemporary computer platforms.

DOI: 10.1055/s-2008-1034876

PMID: 9677549 [Indexed for MEDLINE]

3817. Anal Quant Cytol Histol. 1998 Apr;20(2):127-32.

Microscope remote control with an Internet browser.

Wolf G(1), Petersen I, Dietel M.

Author information:

(1)Institute of Pathology, Charité Medical School, Humboldt University of Berlin,

Germany.

OBJECTIVE: To develop a telemicroscopy system that is independent of specialized

hardware and software for microscope remote control.

DESCRIPTION OF THE SYSTEM: An automatic microscope mounted on a CCD camera was

connected to a computer that fulfills the function of an internet server. Any

internet user can access this server and can control the microscope by use of an

internet browser with Java support. The system can be tested at

http:@amba.rz.charite.hu-berlin.de/telemic. FUNCTION AND APPLICATIONS: The user

can move the microscope stage, change the magnification or execute any other

microscope operations by pressing buttons of the downloaded telemicroscopy client

program. The new microscope images are transferred automatically. Structures of

interest within the images can be highlighted and discussed online with other

telemicroscopy clients. Any internet browser with Java support, like Netscape or

Microsoft Internet Explorer, makes an internet user a telemicroscopy client and

thus a possible consultant for telepathology.

CONCLUSION: The internet telemicroscope offers new possibilities for

telepathology. This development should promote communication between pathologists

and may thus increase the quality of diagnosis.

PMID: 9569970 [Indexed for MEDLINE]

3818. Telemed Today. 1998 Apr-May;6(2):16-8.

From early wireless to Everest.

Allen A.

Medical information has been transmitted using wireless technologies for almost

80 years. A "wired wireless" electronic stethoscope was developed by the U.S.

Army Signal Corps in the early 1920's, for potential use in ship-to-shore

transmission of cardiac sounds. [Winters SR. Diagnosis by wireless. Scientific

American June 11, 1921, p. 465] Today, wireless is used in a wide range of

medical applications and at sites from transoceanic air flights to offshore oil

platforms to Mt. Everest. 'Wireless LANs' are often used in medical environments.

Typically, nurses and physicians in a hospital or clinic use hand-held "wireless

thin client" pen computers that exchange patient information and images with the

hospital server. Numerous companies, such as Fujitsu (article below) and Cruise

Technologies (www.cruisetech.com) manufacture handheld pen-entry computers. One

company, LXE, integrates radio-frequency (RF) enhanced hand-held computers

specifically designed for production use within a wireless LAN (www.lxe.com).

Other companies (Proxim, Symbol, and others) supply the wireless RF LAN

infrastructure for the enterprise. Unfortunately, there have been problems with

widespread deployment of wireless LANs. Perhaps the biggest impediment has been

the lack of standards. Although an international standard (IEEE 802.11) was

adopted in 1997, most wireless LAN products still are not compatible with the

equipment of competing companies. A problem with the current standard for LAN

adapters is that throughput is limited to 3 Mbps--compared to at least 10 Mbps,

and often 100 Mbps, in a hard-wired Ethernet LAN. An II Mbps standard is due out

in the next year or so, but it will be at least 2 years before

standards-compliant products are available. This story profiles some of the ways

that wireless is being used to overcome gaps in terrestrial and within-enterprise

communication.

PMID: 10181174 [Indexed for MEDLINE]

3819. Hemoglobin. 1998 Mar;22(2):113-27.

Access to a syllabus of human hemoglobin variants (1996) via the World Wide Web.

Hardison RC(1), Chui DH, Riemer CR, Miller W, Carver MF, Molchanova TP, Efremov

GD, Huisman TH.

Author information:

(1)Department of Biochemistry and Molecular Biology, Center for Gene Regulation,

The Pennsylvania State University, University Park 16802, USA. rch8@psu

Information on mutations in human hemoglobin is important in many efforts,

including understanding the pathophysiology of hemoglobin diseases, developing

therapies, elucidating the dynamics of sequence alterations inhuman populations,

and dissecting the details of protein structure/function relationships.

Currently, information is available on a large number of mutations and variants,

but is distributed among thousands of papers. In an effort to organize this

voluminous data set, two Syllabi have been prepared compiling succinct

information on human hemoglobin abnormalities. In both of these, each entry

provides amino acid and/or DNA sequence alterations, hematological and clinical

data, methodology used for characterization, ethnic distribution, and functional

properties and stability of the hemoglobin, together with appropriate literature

references. A Syllabus of Human Hemoglobin Variants (1996) describes 693 abnormal

hemoglobins resulting from alterations in the alpha-, beta-, gamma-, and

delta-globin chains, including special abnormalities such as double mutations,

hybrid chains, elongated chains, deletions, and insertions. We have converted

this resource to an electronic form that is accessible via the World Wide Web at

the Globin Gene Server (http://globin.cse.psu.edu). Hyperlinks are provided from

each entry in the tables of variants to the corresponding full description. In

addition, a simple query interface allows the user to find all entries containing

a designated word or phrase. We are in the process of converting A Syllabus of

Thalassemia Mutations (1997) to a similar electronic format.

PMID: 9576329 [Indexed for MEDLINE]

3820. Br J Radiol. 1998 Feb;71(842):167-74.

Distance learning in mammographic digital image processing.

Costaridou L(1), Panayiotakis G, Sakellaropoulos P, Cavouras D, Dimopoulos J.

Author information:

(1)Department of Medical Physics, School of Medicine, University of Patras,

Greece.

The potential of interactive multimedia and Internet technologies is investigated

with respect to the implementation of a distance learning system in medical

imaging. The system is built according to a client-server architecture, based on

the Internet infrastructure, composed of server nodes conceptually modelled as

World Wide Web (WWW) sites. Sites are implemented by integration and

customization of available components. The system evolves around

network-delivered interactive multimedia courses and network-based tutoring,

which constitute its main learning features. This potential has been demonstrated

by means of an implemented system, validated with digital image processing

content, specifically image enhancement. Image enhancement methods are

theoretically described and applied on mammograms. Emphasis is given in the

interactive presentation of the effects of algorithm parameters on images. The

system end-user access depends on available bandwidth, so high speed access can

be achieved via LAN or local ISDN connections.

DOI: 10.1259/bjr.71.842.9579181

PMID: 9579181 [Indexed for MEDLINE]

3821. Glycoconj J. 1998 Feb;15(2):115-30.

NetOglyc: prediction of mucin type O-glycosylation sites based on sequence

context and surface accessibility.

Hansen JE(1), Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S.

Author information:

(1)Center for Biological Sequence Analysis, The Technical University of Denmark,

Lyngby. janhan@cbs.dtu.dk

The specificities of the UDP-GalNAc:polypeptide Nacetylgalactosaminyltransferases

which link the carbohydrate GalNAc to the side-chain of certain serine and

threonine residues in mucin type glycoproteins, are presently unknown. The

specificity seems to be modulated by sequence context, secondary structure and

surface accessibility. The sequence context of glycosylated threonines was found

to differ from that of serine, and the sites were found to cluster. Non-clustered

sites had a sequence context different from that of clustered sites. Charged

residues were disfavoured at position -1 and +3. A jury of artificial neural

networks was trained to recognize the sequence context and surface accessibility

of 299 known and verified mucin type O-glycosylation sites extracted from

O-GLYCBASE. The cross-validated NetOglyc network system correctly found 83% of

the glycosylated and 90% of the non-glycosylated serine and threonine residues in

independent test sets, thus proving more accurate than matrix statistics and

vector projection methods. Predictions of O-glycosylation sites in the envelope

glycoprotein gp120 from the primate lentiviruses HIV-1, HIV-2 and SIV are

presented. The most conserved O-glycosylation signals in these

evolutionary-related glycoproteins were found in their first hypervariable loop,

V1. However, the strain variation for HIV-1 gp120 was significant. A computer

server, available through WWW or E-mail, has been developed for prediction of

mucin type O-glycosylation sites in proteins based on the amino acid sequence.

The server addresses are http://www.cbs.dtu.dk/services/NetOGlyc/ and

netOglyc@cbs.dtu.dk.

PMID: 9557871 [Indexed for MEDLINE]

3822. Bioinformatics. 1998;14(10):892-3.

JPred: a consensus secondary structure prediction server.

Cuff JA(1), Clamp ME, Siddiqui AS, Finlay M, Barton GJ.

Author information:

(1)Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road,

Oxford OX1 3QU, UK.

An interactive protein secondary structure prediction Internet server is

presented. The server allows a single sequence or multiple alignment to be

submitted, and returns predictions from six secondary structure prediction

algorithms that exploit evolutionary information from multiple sequences. A

consensus prediction is also returned which improves the average Q3 accuracy of

prediction by 1% to 72.9%. The server simplifies the use of current prediction

algorithms and allows conservation patterns important to structure and function

to be identified.AVAILABILITY: http://barton.ebi.ac.uk/servers/jpred.h tml

CONTACT: geoff@ebi.ac.uk

PMID: 9927721 [Indexed for MEDLINE]

3823. Bioinformatics. 1998;14(10):846-56.

Hidden Markov models for detecting remote protein homologies.

Karplus K(1), Barrett C, Hughey R.

Author information:

(1)Department of Computer Engineering, Jack Baskin School of Engineering,

University of California, Santa Cruz, CA 95064, USA.

MOTIVATION: A new hidden Markov model method (SAM-T98) for finding remote

homologs of protein sequences is described and evaluated. The method begins with

a single target sequence and iteratively builds a hidden Markov model (HMM) from

the sequence and homologs found using the HMM for database search. SAM-T98 is

also used to construct model libraries automatically from sequences in structural

databases.

METHODS: We evaluate the SAM-T98 method with four datasets. Three of the test

sets are fold-recognition tests, where the correct answers are determined by

structural similarity. The fourth uses a curated database. The method is compared

against WU-BLASTP and against DOUBLE-BLAST, a two-step method similar to ISS, but

using BLAST instead of FASTA.

RESULTS: SAM-T98 had the fewest errors in all tests-dramatically so for the

fold-recognition tests. At the minimum-error point on the SCOP (Structural

Classification of Proteins)-domains test, SAM-T98 got 880 true positives and 68

false positives, DOUBLE-BLAST got 533 true positives with 71 false positives, and

WU-BLASTP got 353 true positives with 24 false positives. The method is optimized

to recognize superfamilies, and would require parameter adjustment to be used to

find family or fold relationships. One key to the performance of the HMM method

is a new score-normalization technique that compares the score to the score with

a reversed model rather than to a uniform null model.

AVAILABILITY: A World Wide Web server, as well as information on obtaining the

Sequence Alignment and Modeling (SAM) software suite, can be found at

http://www.cse.ucsc.edu/research/compbi o/

CONTACT: karplus@cse.ucsc.edu; http://www.cse.ucsc.edu/karplus

PMID: 9927713 [Indexed for MEDLINE]

3824. Bioinformatics. 1998;14(9):827-8.

An Internet linkage and mutation database for the complex phenotype asthma.

Wjst M(1), Immervoll T.

Author information:

(1)GSF National Research Center for Environment and Health, P.O.B. 1129, D-85758

Oberschleissheim, Germany. wjst@gsf.de

SUMMARY: The paper presents details of database construction, website

installation and server architecture of the asthma and allergy gene database.

AVAILABILITY: Database and server templates are available on request from the

first author.

SUPPLEMENTARY INFORMATION: The URL of the asthma and allergy gene database is

http://cooke.gsf.de

PMID: 9918958 [Indexed for MEDLINE]

3825. Bioinformatics. 1998;14(9):772-82.

Weighting hidden Markov models for maximum discrimination.

Karchin R(1), Hughey R.

Author information:

(1)Department of Computer Engineering, Jack Baskin School of Engineering,

University of California, Santa Cruz, CA 95064, USA. rph@cse.ucsc.edu

MOTIVATION: Hidden Markov models can efficiently and automatically build

statistical representations of related sequences. Unfortunately, training sets

are frequently biased toward one subgroup of sequences, leading to an

insufficiently general model. This work evaluates sequence weighting methods

based on the maximum-discrimination idea.

RESULTS: One good method scales sequence weights by an exponential that ranges

between 0.1 for the best scoring sequence and 1.0 for the worst. Experiments with

a curated data set show that while training with one or two sequences performed

worse than single-sequence Probabilistic Smith-Waterman, training with five or

ten sequences reduced errors by 20% and 51%, respectively. This new version of

the SAM HMM suite outperforms HMMer (17% reduction over PSW for 10 training

sequences), Meta-MEME (28% reduction), and unweighted SAM (31% reduction).

AVAILABILITY: A WWW server, as well as information on obtaining the Sequence

Alignment and Modeling (SAM) software suite and additional data from this work,

can be found at http://www.cse.ucse. edu/research/compbio/sam.html

PMID: 9918947 [Indexed for MEDLINE]

3826. Bioinformatics. 1998;14(8):749-50.

A web server to locate periodicities in a sequence

Pasquier CM(1), Promponas VI VI, Varvayannis NJ, Hamodrakas SJ.

Author information:

(1)Department of Biology, Division of Cell Biology and Biophysics, University of

Athens, Athens 15701, Greece.

Summary : FT is a tool written in C++, which implements the Fourier analysis

method to locate periodicities in aminoacid or DNA sequences. It is provided for

free public use on a WWW server with a Java interface. Availability : The server

address is http://o2.db. uoa.gr/FT Contact : shamodr@atlas.uoa.gr

PMID: 9789101

3827. Bioinformatics. 1998;14(8):665-75.

Gene recognition by combination of several gene-finding programs.

Murakami K(1), Takagi T.

Author information:

(1)1Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1

Shirokane-dai Minato-ku, Tokyo 108-8639 and 2Central Research Laboratory, Hitachi

Ltd, 1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan.

MOTIVATION: A number of programs have been developed to predict the eukaryotic

gene structures in DNA sequences. However, gene finding is still a challenging

problem.

RESULTS: We have explored the effectiveness when the results of several

gene-finding programs were re-analyzed and combined. We studied several methods

with four programs (FEXH, GeneParser3, GEN-SCAN and GRAIL2). By HIGHEST-policy

combination method or BOUNDARY method, approximate correlation (AC) improved by

3-5% in comparison with the best single gene-finding program. From another

viewpoint, OR-based combination of the four programs is the most reliable to know

whether a candidate exon overlaps with the real exon or not, although it is less

sensitive than GENSCAN for exon-intron boundaries. Our methods can easily be

extended to combine other programs.

AVAILABILITY: We have developed a server program (Shirokane System) and a client

program (GeneScope) to use the methods. GeneScope is available through a WWW site

(http://gf.genome.ad.jp/).

CONTACT: (katsu,takagi)@ims.u-tokyo.ac.jp

PMID: 9789092 [Indexed for MEDLINE]

3828. Bioinformatics. 1998;14(7):617-23.

JOY: protein sequence-structure representation and analysis.

Mizuguchi K(1), Deane CM, Blundell TL, Johnson MS, Overington JP.

Author information:

(1)Department of Biochemistry, University of Cambridge, Tennis Court Road,

Cambridge CB2 1GA, UK. kenji@cryst.bioc.cam.ac.uk

MOTIVATION: JOY is a program to annotate protein sequence alignments with

three-dimensional (3D) structural features. It was developed to display 3D

structural information in a sequence alignment and to help understand the

conservation of amino acids in their specific local environments.

RESULTS: : The JOY representation now constitutes an essential part of the two

databases of protein structure alignments: HOMSTRAD

(http://www-cryst.bioc.cam.ac.uk/homstrad ) and CAMPASS

(http://www-cryst.bioc.cam.ac. uk/campass). It has also been successfully used

for identifying distant evolutionary relationships.

AVAILABILITY: The program can be obtained via anonymous ftp from

torsa.bioc.cam.ac.uk from the directory /pub/joy/. The address for the JOY server

is http://www-cryst.bioc.cam.ac.uk/cgi-bin/joy.cgi.

CONTACT: kenji@cryst.bioc.cam.ac.uk

PMID: 9730927 [Indexed for MEDLINE]

3829. Bioinformatics. 1998;14(7):600-7.

Automatic extraction of keywords from scientific text: application to the

knowledge domain of protein families.

Andrade MA(1), Valencia A.

Author information:

(1)Protein Design Group, CNB-CSIC, Cantoblanco, E-28049 Madrid, Spain.

MOTIVATION: Annotation of the biological function of different protein sequences

is a time-consuming process currently performed by human experts. Genome analysis

tools encounter great difficulty in performing this task. Database curators,

developers of genome analysis tools and biologists in general could benefit from

access to tools able to suggest functional annotations and facilitate access to

functional information.

APPROACH: We present here the first prototype of a system for the automatic

annotation of protein function. The system is triggered by collections of s

related to a given protein, and it is able to extract biological information

directly from scientific literature, i.e. MEDLINE abstracts. Relevant keywords

are selected by their relative accumulation in comparison with a domain-specific

background distribution. Simultaneously, the most representative sentences and

MEDLINE abstracts are selected and presented to the end-user. Evolutionary

information is considered as a predominant characteristic in the domain of

protein function. Our system consequently extracts domain-specific information

from the analysis of a set of protein families.

RESULTS: The system has been tested with different protein families, of which

three examples are discussed in detail here: 'ataxia-telangiectasia associated

protein', 'ran GTPase' and 'carbonic anhydrase'. We found generally good

correlation between the amount of information provided to the system and the

quality of the annotations. Finally, the current limitations and future

developments of the system are discussed.

AVAILABILITY: The current system can be considered as a prototype system. As

such, it can be accessed as a server at http://columba.ebi.ac.

uk:8765/andrade/abx. The system accepts text related to the protein or proteins

to be evaluated (optimally, the result of a MEDLINE search by keyword) and the

results are returned in the form of Web pages for keywords, sentences and s.

SUPPLEMENTARY INFORMATION: Web pages containing full information on the examples

mentioned in the text are available at: http://www.cnb.uam.es/ approximately

cnbprot/keywords/

CONTACT: valencia@cnb.uam.es

PMID: 9730925 [Indexed for MEDLINE]

3830. Bioinformatics. 1998;14(6):523-8.

Homology modeling, model and software evaluation: three related resources.

Rodriguez R(1), Chinea G, Lopez N, Pons T, Vriend G.

Author information:

(1)Center for Genetic Engineering and Biotechnology, Havana, Cuba and

2BIOcomputing, EMBL, Heidelberg, Germany.

MOTIVATION: Homology modeling is rapidly becoming the method of choice for

obtaining three-dimensional coordinates for proteins because genome projects

produce sequences at a much higher rate than NMR and X-ray laboratories can solve

the three-dimensional structures. The quality of protein models will not be

immediately clear to novices and support with the evaluation seems to be needed.

Expert users are sometimes interested in evaluating the quality of modeling

programs rather than the quality of the models themselves.

RESULTS: Three servers have been made available to the scientific community: a

homology modeling server, a model quality evaluation server and a server that

evaluates models built of proteins for which the structure is already known,

thereby implicitly evaluating the quality of the modeling program.

AVAILABILITY: The modeling-related servers and several structure analysis servers

are freely available at http://swift.embl-heidelberg.de/servers/

CONTACT: gert.vriend@embl-heidelberg.de

PMID: 9694991 [Indexed for MEDLINE]

3831. Bioinformatics. 1998;14(4):342-8.

The new Virgil database: a service of rich links.

Achard F(1), Cussat-Blanc C, Viara E, Barillot E.

Author information:

(1)GIS Infobiogen, 7 rue Guy Môquet, 94801 Villejuif, France and Sysra

informatique, 7 rue de Bievres, 92140 Clamart, France.

Frederic.Achard@infobiogen.fr

MOTIVATION: Links between biological objects are frequently used by researchers

in biology. However, many of the links found in public databases are

insufficiently documented and difficult to retrieve. Virgil introduces the idea

of a rich link, i.e. the link itself and the related pieces of information.

Virgil was developed to collect, manage and distribute such links.

RESULTS: At the moment, Virgil is a prototype database that contains rich links

between GDB genes and Genbank sequences. The Virgil data model is rich enough to

describe comprehensively a link between two biological objects. Two different

means to access the information were developed: a schema-driven Web interface and

a CORBA server.

AVAILABILITY: http://www.infobiogen. fr/services/virgil/home.html

CONTACT: Frederic.Achard@infobiogen.fr

PMID: 9632829 [Indexed for MEDLINE]

3832. Bioinformatics. 1998;14(3):252-8.

A distributed environment for physical map construction.

Grigoriev A(1), Levin A, Lehrach H.

Author information:

(1)Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin,

Germany.

MOTIVATION: With the main focus of the Human Genome Project shifting to

sequencing, bioinformatics support for constructing large-scale genomic maps of

other organisms is still required. We attempt to provide for this with our work,

aimed at the delivery of robust and user-friendly contig-building software on the

WWW.

RESULTS: We present a prototype distributed analytical environment for molecular

biologists working in the area of genomic mapping. It consists of the WWW server

for constructing contigs from users' data with a hypertext output connected to

Java-based map visualization software.

AVAILABILITY: Freely available on http://www.mpimg-berlin-dahlem.mpg. de/

approximately andy/server/

CONTACT: andy@rag3.rz-berlin.mpg.de

PMID: 9614268 [Indexed for MEDLINE]

3833. Bioinformatics. 1998;14(2):223-4.

GeneFIND web server for protein family identification and information retrieval.

Wu CH(1), Shivakumar S, Shivakumar CV, Chen SC.

Author information:

(1)Department of Epidemiology/Biomathematics, The University of Texas Health

Center at Tyler 75710, USA.

An integrated database and search system has been developed for protein family

identification and information retrieval, as an approach to undertake the highly

complex, genomic-scale problem of molecular sequence database search and

organization.AVAILABILITY: http://diana.uthct.edu

CONTACT: wu@uthct.edu

PMID: 9545458 [Indexed for MEDLINE]

3834. Health Care Strateg Manage. 1998 Jan;16(1):2-3.

The Health Care News Server reports business, political and managed care news at

www.healthcarenewsserver.com.

Johnson DE.

PMID: 10175800 [Indexed for MEDLINE]

3835. J Automat Chem. 1998;20(3):77-81. doi: 10.1155/S1463924698000091.

LIS-lnterlink-connecting laboratory information systems to remote primary

health-care centres via the Internet.

Clark B(1), Wachowiak B, Crawford EW, Jakubowski Z, Kabata J.

Author information:

(1)Glasgow University Department of Pathological Biochemistry Gartnavel General

Hospital 1053 Gt Western Rd, Glasgow G12 0YN UK.

A pilot study was performed to evaluate the feasibility of using the Internet to

securely deliver patient laboratory results, and the system has subsequently gone

into routine use in Poland. The system went from design to pilot and then to live

implementation within a four-month period, resulting in the LIS-Interlink

software product. Test results are retrieved at regular intervals from the

BioLink(TM) LIS (Laboratory Information System), encrypted and transferred to a

secure area on the Web server. The primary health-care centres dial into the

Internet using a local-cell service provided by Polish Telecom (TP), obtain a

TCP/IP address using the TP DHCP server, and perform HTTP 'get' and 'post'

operations to obtain the files by secure handshaking. The data are then

automatically inserted into a local SQL database (with optional printing of

incoming reports)for cumulative reporting and searching functions. The local

database is fully multi-user and can be accessed from different clinics within

the centres by a variety of networking protocols.

DOI: 10.1155/S1463924698000091

PMCID: PMC2548147

PMID: 18924820

3836. J Telemed Telecare. 1998;4 Suppl 1:93-4.

Easy Medic: an Internet application for the general practitioner.

Arnone G(1), Bianchi A, Della Pietra B, Sernicola R, Sparacino E, Vitolo R.

Author information:

(1)Consorzio Corinto, Naples, Italy.

A research project has been carried out to develop a client server application

which supplies the general practitioner (GP) with a 'personal digital assistant'

(hand-held mobile computer) to connect to Web servers at a hospital site through

the Internet. This allows the doctor to book medical examinations, hospital

admissions and manage patient data. The application used advanced object-oriented

techniques, on both the client and the server side. The connection to a Web

server was achieved through GSM wireless cellular telephones using standard

Internet protocols (HTTP, TCP/IP and CGI). Conventional telephone lines can be

used as well. Other application modules on the client side provided patients

medical record supervision, GP schedule management, general information about

hospitals and clinics, and pharmacy consultation. These services should help GPs

in their daily work. Moreover, the quality of health-care resource management and

cost supervision should improve, since each GP 'transaction' is automatically

entered in realtime into a database at the server. The services are under test in

the health-care system of an urban area in southern Italy.

DOI: 10.1258/1357633981931623

PMID: 9640752 [Indexed for MEDLINE]

3837. Nihon Igaku Hoshasen Gakkai Zasshi. 1998 Jan;58(1):34-7.

[Management and delivery of radiation dose distribution images using the

Internet].

[Article in Japanese]

Onogi Y(1), Nakagawa K, Aoki Y, Kozuka T, Toyoda T, Sasaki Y.

Author information:

(1)Department of Radiology, University of Tokyo Hospital.

Dose distribution images play important roles in the management of cancer

patients. To date hard copies of these images have been stored and referred to by

radiation oncologists as needed. In most cases, these images were not available

to medical personnel outside the radiation oncology department. We have developed

a mechanism in the hospital to access these dose distribution images via WWW

(World Wide Web). A screen snapshot of a dose distribution image on the CRT of a

treatment planning machine is copied to the WWW server and converted to a GIF

image. Similarly, we can register dose volume histograms and digitally

reconstructed radiographs on the WWW. Medical personnel throughout the hospital

can access the images through the WWW browser. As a result, radiation oncologists

are given detailed information on target definition in treatment planning by

expert physicians. The system also helps co-medical staff in understanding dose

distributions and predicting radiation injuries. At the same time, it actualizes

an electronic archive of dose distribution images, which is a database for quick

and reliable review, evaluation and comparison of treatment plans. This technique

also furthers a close relationship among radiation oncologists, physicians, and

co-medical personnel.

PMID: 9493431 [Indexed for MEDLINE]

3838. Nucleic Acids Res. 1998 Jan 1;26(1):323-6.

The ProDom database of protein domain families.

Corpet F(1), Gouzy J, Kahn D.

Author information:

(1)Laboratoire de Génétique Cellulaire and Laboratoire de Biologie Moléculaire

des Relations Plantes-Microorganismes, INRA/CNRS, BP 27, F-31326 Castanet-Tolosan

Cedex, France.

The ProDom database contains protein domain families generated from the

SWISS-PROT database by automated sequence comparisons. It can be searched on the

World Wide Web (http://protein.toulouse.inra. fr/prodom.html ) or by E-mail

(prodom@toulouse.inra.fr) to study domain arrangements within known families or

new proteins. Strong emphasis has been put on the graphical user interface which

allows for interactive analysis of protein homology relationships. Recent

improvements to the server include: ProDom search by keyword; links to PROSITE

and PDB entries; more sensitive ProDom similarity search with BLAST or WU-BLAST;

alignments of query sequences with homologous ProDom domain families; and links

to the SWISS-MODEL server (http: //www.expasy.ch/swissmod/SWISS-MODEL.html ) for

homology based 3-D domain modelling where possible.

PMCID: PMC147246

PMID: 9399865 [Indexed for MEDLINE]

3839. Nucleic Acids Res. 1998 Jan 1;26(1):309-12.

Superior performance in protein homology detection with the Blocks Database

servers.

Henikoff S(1), Pietrokovski S, Henikoff JG.

Author information:

(1)Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, 1100

Fairview Avenue North, Seattle, WA 98109-1024, USA. steveh@muller.fhcrc.org

The Blocks Database World Wide Web (http://www.blocks.fhcrc.org ) and Email

(blocks@blocks.fhcrc.org) servers provide tools for the detection and analysis of

protein homology based on alignment blocks representing conserved regions of

proteins. During the past year, searching has been augmented by supplementation

of the Blocks Database with blocks from the Prints Database, for a total of 4754

blocks from 1163 families. Blocks from both the Blocks and Prints Databases and

blocks that are constructed from sequences submitted to Block Maker can be used

for blocks-versus-blocks searching of these databases with LAMA, and for viewing

logos and bootstrap trees. Sensitive searches of up-to-date protein sequence

databanks are carried out via direct links to the MAST server using

position-specific scoring matrices and to the BLAST and PSI-BLAST servers using

consensus-embedded sequence queries. Utilizing the trypsin family to evaluate

performance, we illustrate the superiority of blocks-based tools over expert

pairwise searching or Hidden Markov Models.

PMCID: PMC147168

PMID: 9399861 [Indexed for MEDLINE]

3840. Nucleic Acids Res. 1998 Jan 1;26(1):304-8.

The PRINTS protein fingerprint database in its fifth year.

Attwood TK(1), Beck ME, Flower DR, Scordis P, Selley JN.

Author information:

(1)Department of Biochemistry and Molecular Biology, University College London,

London WCIE 6BT, UK. attwood@biochemistry.ucl.ac.uk

PRINTS is a database of protein family 'fingerprints' offering a diagnostic

resource for newly-determined sequences. By contrast with PROSITE, which uses

single consensus expressions to characterise particular families, PRINTS exploits

groups of motifs to build characteristic signatures. These signatures offer

improved diagnostic reliability by virtue of the mutual context provided by motif

neighbours. To date, 800 fingerprints have been constructed and stored in PRINTS.

The current version, 17.0, encodes approximately 4500 motifs, covering a range of

globular and membrane proteins, modular polypeptides, and so on. The database is

accessible via the UCL Bioinformatics World Wide Web (WWW) Server at http://www.

biochem.ucl.ac.uk/bsm/dbbrowser/ . We have recently enhanced the usefulness of

PRINTS by making available new, intuitive search software. This allows both

individual query sequence and bulk data submission, permitting easy analysis of

single sequences or complete genomes. Preliminary results indicate that use of

the PRINTS system is able to assign additional functions not found by other

methods, and hence offers a useful adjunct to current genome analysis protocols.

PMCID: PMC147187

PMID: 9399860 [Indexed for MEDLINE]

3841. Nucleic Acids Res. 1998 Jan 1;26(1):226-8.

aCHEdb: the database system for ESTHER, the alpha/beta fold family of proteins

and the Cholinesterase gene server.

Cousin X(1), Hotelier T, Giles K, Toutant JP, Chatonnet A.

Author information:

(1)Différenciation Cellulaire et Croissance and Unité Informatique, INRA-ENSAM, 2

place Viala, 34060 Montpellier, France.

Acetylcholinesterase belongs to a family of proteins, the alpha/beta hydrolase

fold family, whose constituents evolutionarily diverged from a common ancestor

and share a similar structure of a central beta sheet surrounded by alpha

helices. These proteins fulfil a wide range of physiological functions

(hydrolases, adhesion molecules, hormone precursors) [Krejci,E., Duval,N.,

Chatonnet,A., Vincens,P. and Massoulié,J. (1991) Proc. Natl. Acad. Sci. USA , 88,

6647-6651]. ESTHER (for esterases, alpha/beta hydrolase enzymes and relatives) is

a database aimed at collecting in one information system, sequence data together

with biological annotations and experimental biochemical results related to the

structure-function analysis of the enzymes of the family. The major upgrade of

the database comes from the use of a new database management system: aCHEdb which

uses the ACeDB program designed by Richard Durbin and Jean Thierry-Mieg. It can

be found at http://www.ensam.inra.fr/cholinesterase

PMCID: PMC147245

PMID: 9399841 [Indexed for MEDLINE]

3842. Nucleic Acids Res. 1998 Jan 1;26(1):205-13.

IARC Database of p53 gene mutations in human tumors and cell lines: updated

compilation, revised formats and new visualisation tools.

Hainaut P(1), Hernandez T, Robinson A, Rodriguez-Tome P, Flores T, Hollstein M,

Harris CC, Montesano R.

Author information:

(1)International Agency for Research on Cancer, 150 cours Albert-Thomas, 69372

Lyon Cedex 08, France. hainaut@iarc.fr

Since 1989, about 570 different p53 mutations have been identified in more than

8000 human cancers. A database of these mutations was initiated by M. Hollstein

and C. C. Harris in 1990. This database originally consisted of a list of somatic

point mutations in the p 53 gene of human tumors and cell lines, compiled from

the published literature and made available in a standard electronic form. The

database is maintained at the International Agency for Research on Cancer (IARC)

and updated versions are released twice a year (January and July). The current

version (July 1997) contains records on 6800 published mutations and will surpass

the 8000 mark in the January 1998 release. The database now contains information

on somatic and germline mutations in a new format to facilitate data retrieval.

In addition, new tools are constructed to improve data analysis, such as a

Mutation Viewer Java applet developed at the European Bioinformatics Institute

(EBI) to visualise the location and impact of mutations on p53 protein structure.

The database is available in different electronic formats at IARC

(http://www.iarc. fr/p53/homepage.htm ) or from the EBI server

(http://www.ebi.ac.uk ). The IARC p53 website also provides reports on database

analysis and links with other p53 sites as well as with related databases. In

this report, we describe the criteria for inclusion of data, the revised format

and the new visualisation tools. We also briefly discuss the relevance of p 53

mutations to clinical and biological questions.

PMCID: PMC147235

PMID: 9399837 [Indexed for MEDLINE]

3843. Nucleic Acids Res. 1998 Jan 1;26(1):183-6.

Database on the structure of large ribosomal subunit RNA.

De Rijk P(1), Caers A, Van de Peer Y, De Wachter R.

Author information:

(1)Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1,

B-2610 Antwerpen, Belgium.

The rRNA WWW Server at URL http://rrna.uia.ac.be/ now provides a database of 496

large subunit ribosomal RNA sequences. All these sequences are aligned,

incorporate secondary structure information, and can be obtained in a number of

formats. Other information about the sequences, such as literature references,

accession numbers and taxonomic information is also available and searchable. If

necessary, the data on the server can also be obtained by anonymous ftp.

PMCID: PMC147202

PMID: 9399830 [Indexed for MEDLINE]

3844. Nucleic Acids Res. 1998 Jan 1;26(1):106-7.

HuGeMap: a distributed and integrated Human Genome Map database.

Barillot E(1), Guyon F, Cussat-Blanc C, Viara E, Vaysseix G.

Author information:

(1)GIS Infobiogen, 7 rue Guy Môquet, BP 8, 94801 Villejuif Cedex, France and

Sysra Informatique, 7 rue de Bièvres, 92140 Clamart, France.

emmanuel.barillot@infobiogen.fr

The HuGeMap database stores the major genetic and physical maps of the human

genome. It is also interconnected with the gene radiation hybrid mapping database

RHdb. HuGeMap is accessible through a Web server for interactive browsing at URL

http://www.infobiogen. fr/services/Hugemap , as well as through a CORBA server

for effective programming. HuGeMap is intended as an attempt to build open,

interconnected databases, that is databases that distribute their objects

worldwide in compliance with a recognized standard of distribution. Maps can be

displayed and compared with a java applet

(http://babbage.infobiogen.fr:15000/Mappet/Show. html ) that queries the HuGeMap

ORB server as well as the RHdb ORB server at the EBI.

PMCID: PMC147238

PMID: 9399811 [Indexed for MEDLINE]

3845. Nucleic Acids Res. 1998 Jan 1;26(1):100-1.

Virgil: a database of rich links between GDB and GenBank.

Achard F(1), Barillot E.

Author information:

(1)GIS Infobiogen, 7 rue Guy Môquet, BP 8, 94801 Villejuif Cedex, France.

frederic.achard@infobiogen.fr

Database interconnection requires the development of links between related

objects from different databases. We built a database of links, called Virgil, to

manage and distribute rich (documented) links between GDB genes and GenBank human

sequences. Virgil contains 18 667 unique links. In addition to a simple Web form

for ad-hoc queries, we propose a generic Web interface and a prototype CORBA

server for link distribution. Materials described in this paper are available

from http://www.infobiogen.fr/services/virgil/home. html

PMCID: PMC147227

PMID: 9399809 [Indexed for MEDLINE]

3846. Nucleic Acids Res. 1998 Jan 1;26(1):33-7.

MIPS: a database for protein sequences and complete genomes.

Mewes HW(1), Hani J, Pfeiffer F, Frishman D.

Author information:

(1)Munich Information Center for Protein Sequences (MIPS/GSF) am

Max-Planck-Institut für Biochemie, Am Klopferspitz 18, D-82152 Martinsried,

Germany. mewes@mips.biochem.mpg.de

The MIPS group [Munich Information Center for Protein Sequences of the German

National Center for Environment and Health (GSF)] at the Max-Planck-Institute for

Biochemistry, Martinsried near Munich, Germany, is involved in a number of data

collection activities, including a comprehensive database of the yeast genome, a

database reflecting the progress in sequencing the Arabidopsis thaliana genome,

the systematic analysis of other small genomes and the collection of protein

sequence data within the framework of the PIR-International Protein Sequence

Database (described elsewhere in this volume). Through its WWW server

(http://www.mips.biochem.mpg.de ) MIPS provides access to a variety of generic

databases, including a database of protein families as well as automatically

generated data by the systematic application of sequence analysis algorithms. The

yeast genome sequence and its related information was also compiled on CD-ROM to

provide dynamic interactive access to the 16 chromosomes of the first eukaryotic

genome unraveled.

PMCID: PMC147239

PMID: 9399795 [Indexed for MEDLINE]

3847. Nucleic Acids Res. 1998 Jan 1;26(1):1-7.

GenBank.

Benson DA(1), Boguski MS, Lipman DJ, Ostell J, Ouellette BF.

Author information:

(1)National Center for Biotechnology Information, National Library of Medicine,

National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD

20894, USA. dab@ncbi.nlm.nih.gov

The GenBank(R) sequence database (http://www.ncbi.nlm.nih.gov/) incorporates DNA

sequences from all available public sources, primarily through the direct

submission of sequence data from individual laboratories and from large-scale

sequencing projects. Most submitters use the BankIt (WWW) or Sequin programs to

send their sequence data. Data exchange with the EMBL Data Library and the DNA

Data Bank of Japan helps ensure comprehensive worldwide coverage. GenBank data is

accessible through NCBI's integrated retrieval system, Entrez , which integrates

data from the major DNA and protein sequence databases along with taxonomy,

genome and protein structure information. MEDLINE(R) abstracts from published

articles describing the sequences are also included as an additional source of

biological annotation. Sequence similarity searching is offered through the BLAST

series of database search programs. In addition to FTP, e-mail and server/client

versions of Entrez and BLAST, NCBI offers a wide range of World Wide Web

retrieval and analysis services of interest to biologists.

PMCID: PMC147205

PMID: 9399790 [Indexed for MEDLINE]

3848. Pac Symp Biocomput. 1998:719-30.

Proclass protein family database: new version with motif alignments.

Wu CH(1), Shivakumar S.

Author information:

(1)Department of Epidemiology/Biomathematics, University of Texas Health Center

at Tyler 75710, USA.

ProClass is a protein family database which organizes non-redundant sequence

entries into families defined collectively by the ProSite patterns and PIR

superfamilies. The database consists of about 100,000 entries, more than half of

which are classified in about 3,000 families. The new version includes links to

various protein family/domain and structural class databases and contains gapped

motif alignments for all ProSite patterns. The motif sequences are retrieved from

both SwissProt and PIR-international databases, including numerous new members

detected by our GeneFIND family identification system. The motif collection

represents a 50% increase from those catalogued in ProSite. The ProClass database

can be used to maximize family information retrieval, help organize protein

sequence databases, and support full-scale genomic annotation. The database and

its query program are freely available for on-line record retrieval and direct

file transfer from our WWW server at http:/(/)diana.uthct.edu/proclass.html+ ++.

PMID: 9697225 [Indexed for MEDLINE]

3849. Pac Symp Biocomput. 1998:683-94.

DBGET/LinkDB: an integrated database retrieval system.

Fujibuchi W(1), Goto S, Migimatsu H, Uchiyama I, Ogiwara A, Akiyama Y, Kanehisa

M.

Author information:

(1)Institute for Chemical Research, Kyoto University, Japan.

The integrated database retrieval system DBGET/LinkDB is the backbone of the

Japanese GenomeNet service. DBGET is used to search and extract entries from a

wide range of molecular biology databases, while LinkDB is used to search and

compute links between entries in different databases. DBGET/LinkDB is designed to

be a network distributed database system with an open architecture, which is

suitable for incorporating local databases or establishing a specialized server

environment. It also has an advantage of simple architecture allowing rapid daily

updates of all the major databases. The WWW version of DBGET/LinkDB at GenomeNet

is integrated with other search tools, such as BLAST, FASTA and MOTIF, and with

local helper applications, such as RasMol. In addition to factual links between

database entries, LinkDB is being extended to included similarity links and

biological links toward computerization of logical reasoning processes.

PMID: 9697222 [Indexed for MEDLINE]

3850. Proc AMIA Symp. 1998:607-11.

DXplain on the Internet.

Barnett GO(1), Famiglietti KT, Kim RJ, Hoffer EP, Feldman MJ.

Author information:

(1)Harvard Medical School, Laboratory of Computer Science, Massachusetts General

Hospital, Boston, USA.

DXplain, a computer-based medical education, reference and decision support

system has been used by thousands of physicians and medical students on

stand-alone systems and over communications networks. For the past two years, we

have made DXplain available over the Internet in order to provide DXplain's

knowledge and analytical capabilities as a resource to other applications within

Massachusetts General Hospital (MGH) and at outside institutions. We describe and

provide the user experience with two different protocols through which users can

access DXplain through the World Wide Web (WWW). The first allows the user to

have direct interaction with all the functionality of DXplain where the MGH

server controls the interaction and the mode of presentation. In the second mode,

the MGH server provides the DXplain functionality as a series of services, which

can be called independently by the user application program.

PMCID: PMC2232149

PMID: 9929291 [Indexed for MEDLINE]

3851. Proc AMIA Symp. 1998:597-601.

MediAgent: a WWW-based scalable and self-learning medical search engine.

Tay J(1), Ke S, Lun KC.

Author information:

(1)Medical Informatics Program, National University of Singapore.

Searching for medical information on the Internet can be tedious and frustrating

due to the number of irrelevant entries returned from generic search engines. We

have developed MediAgent, a scalable search engine that aims to deliver a

web-based medical search solution which is focused, exhaustive and able to keep

improving its databases. The software package can run off a single low-end system

and be scaled into a client-server, distributed computing architecture for

high-end needs. This scalable architecture boosts MediAgent's handling capacity

to tens of millions of web pages. In addition to large volume handling, MediAgent

is designed to be manageable. All subsystems are not only highly configurable,

but also support remote, interactive management and monitoring by the system

administrator.

PMCID: PMC2232218

PMID: 9929289 [Indexed for MEDLINE]

3852. Profiles Healthc Mark. 1998 Jan-Feb;14(1):2, 8.

The Health Care News Server reports business, political and managed care news at

www.HealthCareNewsServer.com.

[No authors listed]

PMID: 10177000 [Indexed for MEDLINE]

3853. Radiographics. 1998 Jan-Feb;18(1):189-94.

SPR online: creating, maintaining, and distributing a virtual professional

society on the Internet.

D'Alessandro MP(1), Galvin JR.

Author information:

(1)Department of Radiology, University of Iowa Hospitals and Clinics, Iowa City

52242-1009, USA.

SPR Online (http:@www.pedrad.org) is a recently developed digital representation

of the Society for Pediatric Radiology (SPR) that enables physicians to access

pertinent information and services on the Internet. SPR Online was organized on

the basis of the five main services of the SPR, which include Administration,

Patient Care, Education, Research, and Meetings. For each service, related

content from the SPR was digitized and placed onto SPR Online. Usage over a

12-month period was evaluated with server log file analysis. A total of 3,209

users accessed SPR Online, viewing 11,246 pages of information. A wide variety of

information was accessed, with that from the Education, Administration, and

Meetings services being the most popular. Fifteen percent of users came from

foreign countries. As a virtual professional society, SPR Online greatly enhances

the power and scope of the SPR and has proved to be a popular resource, meeting

the diverse information needs of an international community of pediatric

radiologists.

DOI: 10.1148/radiographics.18.1.9460116

PMID: 9460116 [Indexed for MEDLINE]

3854. Sb Lek. 1998;99(4):587-91.

[Providing Internet-based information services at the 2nd Medical School of

Charles University].

[Article in Czech]

Vejvalka J(1), Ulrych O, Vorísek M, Mrázek J.

Author information:

(1)Ustav klinické a aplikované informatiky 2. lékarské fakulty Univerzity

Karlovy, Praha, Czech Republic. Jan.Vejvalka@lfmotol.cuni.cz

Development of the internet network at the 2nd Faculty of Medicine, Charles

University from one of the first nodes of Prague Academic Network up to a

complicated infrastructure with many client computers is accompanied with a

corresponding development of information services available in this network. For

today's users, infrastructure of the network is hidden behind the services whose

technical details are not significant. The paper deals with history, current

state and possible perspectives of information services (both basic as e-mail,

WWW and more advanced like specialized database server, proxy, etc.) available to

users at the 2nd Faculty of Medicine--taking into account the development of

information technologies, networking infrastructure and the possibilities and

limitations of co-operation between the faculty and its teaching hospital.

PMID: 10803307 [Indexed for MEDLINE]

3855. Sb Lek. 1998;99(4):583-5.

[The database server for the medical bibliography database at Charles

University].

[Article in Czech]

Vejvalka J(1), Rojíková V, Ulrych O, Vorísek M.

Author information:

(1)Ustav klinické a aplikované informatiky 2. lékarské fakulty, Univerzita

Karlova, Praha, Czech Republic. Jan.Vejvalka@lfmotol.cuni.cz

In the medical community, bibliographic databases are widely accepted as a most

important source of information both for theoretical and clinical disciplines. To

improve access to medical bibliographic databases at Charles University, a

database server (ERL by Silver Platter) was set up at the 2nd Faculty of Medicine

in Prague. The server, accessible by Internet 24 hours/7 days, hosts now 14

years' MEDLINE and 10 years' EMBASE Paediatrics. Two different strategies are

available for connecting to the server: a specialized client program that

communicates over the Internet (suitable for professional searching) and a

web-based access that requires no specialized software (except the WWW browser)

on the client side. The server is now offered to academic community to host

further databases, possibly subscribed by consortia whose individual members

would not subscribe them by themselves.

PMID: 10803306 [Indexed for MEDLINE]

3856. Stud Health Technol Inform. 1998;52 Pt 2:1075-9.

A hospital-wide distributed PACS based on intranet.

Bandon D(1), Ligier Y, Trayser G, Girard C, Logean M, Ratib O.

Author information:

(1)Digital Imaging Unit, University Hospital of Geneva, Switzerland.

David.Bandon@dim.hcuge.ch

A hospital-wide Picture Archiving and Communication System (PACS) is currently

under development at the University Hospital of Geneva. After a first

implementation including two oneterabyte optical libraries, the system is

expanded to integrate all the imaging modalities of the hospital. The new storage

requirement is 10 terabytes to cover three year archive. A large distributed

image archive has been designed including new archive servers for long-term

storage and display servers for medium-term storage. The acquisition, archive and

distribution cycles are performed using separated networks combining Fast

Ethernet and Ethernet. Image files are distributed to the wide-hospital using a

prefetching strategy or an Intranet server, RADIOLAB. The first mode takes

advantage of the fully integrated hospital information system DIOGENE 2 to allow

the automatic retrieval of studies in advance. The second mode provides a

convivial study selection from any conventional WWW (World Wide Web) browser.

Image files are then transmitted to the user's display station using HTTP

(HyperText Transfer Protocol) and handled by OSIRIS software, which acts as a

helper or viewer. Such a system is expected to meet the time requirement, which

is less than three seconds per image.

PMID: 10384626 [Indexed for MEDLINE]

3857. Stud Health Technol Inform. 1998;52 Pt 2:917-21.

On the way to a Web based hospital information system: concepts for the use of a

medical data dictionary to present context sensitive information in an intranet

environment.

Bürkle T(1), Ruan W, Michel A, Dudeck J.

Author information:

(1)Department of Medical Informatics, University of Giessen, Germany.

Many authors have promoted the www-paradigm to build modern hospital information

systems. However currently web-based applications are better suited to

information "browsing" than to build complex data entry features. This prompted

us to start with our first web based developments inside the Giessen University

Hospital Information System in the field of pure presentation of stored knowledge

and information. This article describes the concepts which will be used in

Giessen to convert available information sources to the www paradigm and to

implement context-sensitive knowledge presentation mechanisms inside the clinical

information system. The approach is based upon a web-based medical data

dictionary server. The data dictionary is used to map terms of interest, chosen

from the clinical user during work with a HIS-application, to a semantic network

of relationships. The dictionary server will follow those semantic links in order

to find and display the webpages, which are linked to the subject.

PMID: 10384592 [Indexed for MEDLINE]

3858. Stud Health Technol Inform. 1998;52 Pt 1:351-5.

IMGT, the international ImMunoGeneTics database: a new design for immunogenetics

data access.

Giudicelli V(1), Chaume D, Mennessier G, Althaus HH, Müller W, Bodmer J, Malik A,

Lefranc MP.

Author information:

(1)Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UMR 5535 (CNRS, Université

Montpellier II), France.

IMGT, the international ImMunoGeneTics database is an integrated database

specializing in Immunoglobulins (Ig), T-cell receptors (TcR) and MHC molecules of

all vertebrate species, created by Marie-Paule Lefranc, University of

Montpellier, CNRS, Montpellier, France (Nucleic Acids Research, Database issue,

Vol 26, January 1998). IMGT includes three databases: LIGM-DB (for Ig and TcR),

MHC/HLA-DB and IMGT/PRIMER-DB (an Ig, TcR and MHC-related primer database), the

last two in development. IMGT comprises expertly annotated sequences and

alignment tables. LIGM-DB contains more than 24.000 Immunoglobulin and T cell

Receptor sequences from 81 different species. MHC/HLA-DB contains class I and

class II Human Leucocyte Antigen alignment tables. An IMGT tool, DNAPLOT,

developed for Ig, TcR and MHC sequence analysis, is also available. IMGT goals

are to establish a common data access to all immunogenetics data, including

nucleotide and protein sequences, oligonucleotide primers, gene maps and other

genetic data of Ig, TcR and MHC molecules, from all species, and to provide a

graphical user friendly data access. IMGT has important implications in medical

research (repertoire in autoimmune diseases, AIDS, leukemias, lymphomas),

therapeutical approaches (antibody engineering), genome diversity and genome

evolution studies. In this paper, we describe our approach for the data

modelisation, the automation of the annotation procedure and control of data

quality in LIGM-DB database. IMGT is freely available on the CNUSC WWW server at

Montpellier: http://imgt.cnusc.fr: 8104 (contact: Denys.Chaume@cnusc.fr) and on

the EBI servers: http://www.ebi.ac.uk/imgt (contact: malik@ebi.ac.uk) and

ftp.ebi.ac.uk/pub/databases/imgt. LIGM-DB users are encouraged to report errors

or suggestions to giudi@ligm.crbm.cnrs-mop.fr. IMGT initiator and coordinator:

Marie-Paule Lefranc, lefranc@ligm.crbm.cnrs-mop.fr. (fax: +33(0)467040231).

PMID: 10384476 [Indexed for MEDLINE]

3859. Comput Appl Biosci. 1997 Dec;13(6):619-20.

The Sequence Alerting Server--a new WEB server.

Hegyi H(1), Lai JM, Bork P.

Author information:

(1)EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany.

A Sequence Alerting Server with a WWW interface is described which informs users

with query sequences in database searches about new entries in protein databases

related to their query.AVAILABILITY: The server address is

http://www.bork.embl-heidelberg.de/alerting/.

PMID: 9475991 [Indexed for MEDLINE]

3860. Comput Appl Biosci. 1997 Dec;13(6):609-15.

SAMBA: hardware accelerator for biological sequence comparison.

Guerdoux-Jamet P(1), Lavenier D.

Author information:

(1)IRISA, Rennes, France. lavenier@irisa.fr

MOTIVATION: SAMBA (Systolic Accelerator for Molecular Biological Applications) is

a 128 processor hardware accelerator for speeding up the sequence comparison

process. The short-term objective is to provide a low-cost board to boost PC or

workstation performance on this class of applications. This paper places SAMBA

amongst other existing systems and highlights the original features.

RESULTS: Real performance obtained from the prototype is demonstrated. For

example, a sequence of 300 amino acids is scanned against SWISS-PROT-34 (21 210

389 residues) in 30 s using the Smith and Waterman algorithm. More time-consuming

applications, like the bank-to-bank comparison, are computed in a few hours

instead of days on standard workstations. Technology allows the prototype to fit

onto a single PCI board for plugging into any PC or workstation.

AVAILABILITY: SAMBA can be tested on the WEB server at URL

http://www.irisa.fr/SAMBA/.

PMID: 9475989 [Indexed for MEDLINE]

3861. Comput Appl Biosci. 1997 Dec;13(6):583-6.

Displaying the information contents of structural RNA alignments: the structure

logos.

Gorodkin J(1), Heyer LJ, Brunak S, Stormo GD.

Author information:

(1)Center for Biological Sequence Analysis, Technical University of Denmark,

Lyngby, Denmark. gorodkin@cbs.dtu.dk

MOTIVATION: We extend the standard 'Sequence Logo' method of Schneider and

Stevens (Nucleic Acids Res., 18, 6097-6100, 1990) to incorporate prior

frequencies on the bases, allow for gaps in the alignments, and indicate the

mutual information of base-paired regions in RNA.

RESULTS: Given an alignment of RNA sequences with the base pairings indicated,

the program will calculate the information at each position, including the mutual

information of the base pairs, and display the results in a 'Structure Logo'.

Alignments without base pairing can also be displayed in a 'Sequence Logo', but

still allowing gaps and incorporating prior frequencies if desired.

AVAILABILITY: The code is available from, and an Internet server can be used to

run the program at, http://www.cbs.dtu.dk/gorodkin/appl/slogo. html.

PMID: 9475985 [Indexed for MEDLINE]

3862. Electrophoresis. 1997 Dec;18(15):2759-73.

The 2DWG meta-database of two-dimensional electrophoretic gel images on the

Internet.

Lemkin PF(1).

Author information:

(1)Image Processing Section, NCI-FCRDC/NIH, Frederick, MD 21702, USA.

lemkin@ncifcrf.gov

The 2DWG meta-database is a searchable database of two-dimensional (2-D)

electrophoretic gel images found on the Internet. A meta-database contains

information about locating data in other databases - but not that data itself.

This database was constructed because of a need for an enriched set of World Wide

Web (WWW) locations (URLs) of 2-D gel images on the Internet. These gel images

are used in conjunction with the National Cancer Institute (NCI) Flicker Server

to manipulate and visually compare 2-D gel images across the Internet. User's

gels may also be compared with those in the database. The 2DWG is organized as a

spreadsheet table with each gel image being represented by a row sorted by tissue

type. Data for each gel includes tissue type, species, cell-line, image URL,

database URL, gel protocol, organization URL, image properties, map URL if it

exists, etc. The 2DWG may be searched to find relevant subsets of gels. Searching

is done using the dbEngine - a WWW database search engine which accesses selected

rows of gels from the full 2DWG table. The 2DWG meta-database is accessible on

the WWW at http://www-lecb.ncifcrf.gov/2dwgDB/ and the NCI Flicker server at

http://www-lecb.ncifcrf.gov/flicker/

DOI: 10.1002/elps.1150181510

PMID: 9504808 [Indexed for MEDLINE]

3863. Protein Sci. 1997 Dec;6(12):2628-30.

OLDERADO: on-line database of ensemble representatives and domains. On Line

Database of Ensemble Representatives And DOmains.

Kelley LA(1), Sutcliffe MJ.

Author information:

(1)Department of Chemistry, University of Leicester, United Kingdom.

In cases where the structure of a single protein is represented by an ensemble of

conformations, there is often a need to determine the common features and to

choose a "representative" conformation. This occurs, for example, with structures

determined by NMR spectroscopy, analysis of the trajectory from a molecular

dynamics simulation, or an ensemble of structures produced by comparative

modeling. We reported previously automatic methods for (1) defining the atoms

with low spatial variance across an ensemble (i.e., the "core" atoms) and the

domains in which these atoms lie, and (2) clustering an ensemble into

conformationally related subfamilies. To extend the utility of these methods, we

have developed a freely available server on the World Wide Web at

http:/(/)neon.chem.le.ac.uk/olderado/. This (1) contains an automatically

generated database of representative structures, core atoms, and domains

determined for 449 ensembles of NMR-derived protein structures in the Protein

Data Bank (PDB) in May 1997, and (2) allows the user to upload a PDB-formatted

file containing the coordinates of an ensemble of structures. The server returns

in real time: (1) information on the residues constituting domains: (2) the

structures that constitute each conformational subfamily; and (3) an interactive

java-based three-dimensional viewer to visualise the domains and clusters. Such

information is useful, for example, when selecting conformations to be used in

comparative modeling and when choosing parts of structures to be used in

molecular replacement. Here we describe the OLDERADO server.

DOI: 10.1002/pro.5560061215

PMCID: PMC2143626

PMID: 9416612 [Indexed for MEDLINE]

3864. Rofo. 1997 Dec;167(6):649-51.

[Use of a PACS for ultrasound connection].

[Article in German]

Beissert M(1), Jenett M, Hahn D.

Author information:

(1)Institut für Röntgendiagnostik, Universität Würzburg.

PURPOSE: This report describes the impact of a new digital picture archiving and

communication system-PACS-in a diagnostic ultrasound department. The system

consists of ultrasound workstations, a patient management workstation, a combined

diagnostic workstation/archive server and a WWW client for standard PCs. Data

exchange with other PAC-systems is possible via Ethernet. Staff required

approximately 5 weeks to fully adjust to the new system. Following the adjustment

period, documenting cases digitally reduced our exam times and improved our

availability of images and reports for subsequent examinations, demonstrations,

research and teaching. In the first year after installation, the new ultrasound

PACS has proven its value in daily use.

DOI: 10.1055/s-2007-1015597

PMID: 9465963 [Indexed for MEDLINE]

3865. Int J Med Inform. 1997 Nov;47(1-2):91-9.

The Internet and randomised controlled trials.

Kelly MA(1), Oldham J.

Author information:

(1)Centre for Health Services Research, University of Leeds, UK.

mikek@epid.leeds.ac.uk

Several factors constrain the implementation of Randomised Controlled Trials

(RCTs). To obtain large sample sizes a multicentred multinational trial may be

necessary or a long sampling period. The larger the trial the larger is the unit

cost. To allow larger sample sizes, shorter sampling periods and lower unit

costs, new methods are needed. The Internet and in particular the WWW provides

such an opportunity. The WWW can provide global access, fast interaction and

automation. A prototype Internet Trials Service (ITS) is currently being tested

with a real international clinical trial (the Growth Restriction Intervention

Trial--GRIT). The ITS is hosted on a Web server. It provides a series of HTML

documents that describe the GRIT protocol. Registered centres may enter patients

into the GRIT trial via ITS. Java applets are used to collect trial data before

returning the study number and randomisation. ITS assumes all trial data will be

intercepted by a sniffer. Therefore no information is sent that could

specifically identify a patient, this must be sent later by more secure means.

ITS assumes that trial centres can be spoofed. To authenticate the patients

entered into the trial and the trial data sent, a regular audit report is sent to

each centre by secure means for confirmation. By using Java, a full functional

data entry system can be developed that runs locally within any Java enabled

browser. It can perform data validation locally and also provide a sophisticated

user interface.

PMID: 9506401 [Indexed for MEDLINE]

3866. Mem Inst Oswaldo Cruz. 1997 Nov-Dec;92(6):805-9.

TcruziDB, an integrated database, and the WWW information server for the

Trypanosoma cruzi genome project.

Degrave W(1), de Miranda AB, Amorim A, Brandão A, Aslett M, Vandeyar M.

Author information:

(1)Departamento de Bioquímica e Biologia Molecular, Instituto Oswaldo Cruz, Rio

de Janeiro, Brasil. wdegrave@gene.dbbm.fiocruz.br

Data analysis, presentation and distribution is of utmost importance to a genome

project. A public domain software, ACeDB, has been chosen as the common basis for

parasite genome databases, and a first release of TcruziDB, the Trypanosoma cruzi

genome database, is available by ftp from

ftp://iris.dbbm.fiocruz.br/pub/genomedb/Tcr uziDB as well as versions of the

software for different operating systems

(ftp://iris.dbbm.fiocruz.br/pub/unixsoft/). Moreover, data originated from the

project are available from the WWW server at http://www.dbbm.fiocruz.br. It

contains biological and parasitological data on CL Brener, its karyotype, all

available T. cruzi sequences from Genbank, data on the EST-sequencing project and

on available libraries, a T. cruzi codon table and a listing of activities and

participating groups in the genome project, as well as meeting reports. T. cruzi

discussion lists (tcruzil@iris.dbbm.fiocruz.br and tcgenics@iris.dbbm.fiocruz.br)

are being maintained for communication and to promote collaboration in the genome

project.

PMID: 9580490 [Indexed for MEDLINE]

3867. Proc Natl Acad Sci U S A. 1997 Oct 28;94(22):11929-34.

Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium.

Fischer D(1), Eisenberg D.

Author information:

(1)University of California, Los Angeles-Department of Energy Laboratory of

Structural Biology and Molecular Medicine, Molecular Biology Institute,

University of California, Los Angeles, Box 951570, Los Angeles, CA 90095-1570,

USA.

A crucial step in exploiting the information inherent in genome sequences is to

assign to each protein sequence its three-dimensional fold and biological

function. Here we describe fold assignment for the proteins encoded by the small

genome of Mycoplasma genitalium. The assignment was carried out by our computer

server (http://www.doe-mbi.ucla.edu/people/frsvr/ frsvr. html), which assigns

folds to amino acid sequences by comparing sequence-derived predictions with

known structures. Of the total of 468 protein ORFs, 103 (22%) can be assigned a

known protein fold with high confidence, as cross-validated with tests on known

structures. Of these sequences, 75 (16%) show enough sequence similarity to

proteins of known structure that they can also be detected by traditional

sequence-sequence comparison methods. That is, the difference of 28 sequences

(6%) are assignable by the sequence-structure method of the server but not by

current sequence-sequence methods. Of the remaining 78% of sequences in the

genome, 18% belong to membrane proteins and the remaining 60% cannot be assigned

either because these sequences correspond to no presently known fold or because

of insensitivity of the method. At the current rate of determination of new folds

by x-ray and NMR methods, extrapolation suggests that folds will be assigned to

most soluble proteins in the next decade.

PMCID: PMC23659

PMID: 9342339 [Indexed for MEDLINE]

3868. Analyst. 1997 Oct;122(10):1001-6.

Data analysis using the Internet: the World Wide Web scanning probe microscopy

data analysis system.

Williams PM(1), Davies MC, Roberts CJ, Tendler SJ.

Author information:

(1)Department of Pharmaceutical Sciences, University of Nottingham, UK.

The first interactive world-wide web-based image analysis system is presented

(http://pharm6.pharm.nottingham.ac.uk/processing/main. html). The system,

currently tailored to scanning probe microscopy image data, has been developed to

permit the use of software algorithms developed within our laboratory by

researchers throughout the world. The implementation and functionality of the

scanning probe microscopy server is described. Feedback from users of the

facility has demonstrated its value within the research community, and

highlighted key operational issues which are to be addressed. A future role of

Internet-based data processing software is also discussed.

PMID: 9463946 [Indexed for MEDLINE]

3869. Comput Appl Biosci. 1997 Oct;13(5):545-7.

Seqalert--a daily sequence alertness server for the EMBL and SWISSPROT databases.

Shomer B(1).

Author information:

(1)EMBL Outstation, European Bioinformatics Institute, Hinxton, Cambridge, UK.

bshomer@ebi.ac.uk

MOTIVATION: The aims were to: enable users to deposit complex search profiles

against the sequence databases; interface to an independent Sequence Retrieval

System (SRS) server through the network to perform these searches on a daily

basis through the last day's updates of these databases; mail users the

reformatted search results, enabling local usage when loaded by a WWW browser.

RESULTS: The deposition of one to many search profiles by the user leads to a

daily search of the EMBL and SWISSPROT databases. The search profile is

restricted to entries that were deposited during the last 24 h by using the SRS

query manager to combine search sets. If the search is successful, the resulting

html page is modified from relative URLs to absolute ones, enabling local usage

by loading from disk. The results are sent to the user by e-mail.

PMID: 9367127 [Indexed for MEDLINE]

3870. Int J Neural Syst. 1997 Oct-Dec;8(5-6):581-99.

A neural network method for identification of prokaryotic and eukaryotic signal

peptides and prediction of their cleavage sites.

Nielsen H(1), Engelbrecht J, Brunak S, von Heijne G.

Author information:

(1)Department of Biotechnology, The Technical University of Denmark, Lyngby.

We have developed a new method for the identification of signal peptides and

their cleavage sites based on neural networks trained on separate sets of

prokaryotic and eukaryotic sequences. The method performs significantly better

than previous prediction schemes, and can easily be applied to genome-wide data

sets. Discrimination between cleaved signal peptides and uncleaved N-terminal

signal-anchor sequences is also possible, though with lower precision.

Predictions can be made on a publicly available WWW server:

http://www.cbs.dtu.dk/services/SignalP/.

PMID: 10065837 [Indexed for MEDLINE]

3871. Proteins. 1997 Oct;29(2):252-7.

SAmBA: an interactive software for optimizing the design of biological

macromolecules crystallization experiments.

Audic S(1), Lopez F, Claverie JM, Poirot O, Abergel C.

Author information:

(1)Information Génétique et Structurale, E.P. 91-C.N.R.S., Institut de Biologie

Structurale et Microbiologie, Marseille, France.

SAmBA is a new software for the design of minimal experimental protocols using

the notion of orthogonal arrays of strength 2. The main application of SAmBA is

the search of protein crystallization conditions. Given a user input defining the

relevant effectors/variables (e.g., pH, temperature, salts) and states (e.g., pH:

5, 6, 7 and 8), this software proposes an optimal set of experiments in which all

tested variables and the pairwise interactions between them are symmetrically

sampled. No a priori restrictions on the number and range of experimental

variables is imposed. SAmBA consists of two complementary programs, SAm and BA,

using a simulated annealing approach and a backtracking algorithm, respectively.

The software is freely available as C code or as an interactive JAVA applet at

http:/(/)igs-server.cnrs-mrs.fr.

PMID: 9329089 [Indexed for MEDLINE]

3872. Int J Med Inform. 1997 Aug;46(1):41-51.

A network of web multimedia medical information servers for a medical school and

university hospital.

Denier P(1), Le Beux P, Delamarre D, Fresnel A, Cleret M, Courtin C, Seka LP,

Pouliquen B, Cleran L, Riou C, Burgun A, Jarno P, Leduff F, Lesaux H,

Duvauferrier R.

Author information:

(1)Centre Hospitalier, Rennes, France.

Modern medicine requires a rapid access to information including clinical data

from medical records, bibliographic databases, knowledge bases and nomenclature

databases. This is especially true for University Hospitals and Medical Schools

for training as well as for fundamental and clinical research for diagnosis and

therapeutic purposes. This implies the development of local, national and

international cooperation which can be enhanced via the use and access to

computer networks such as Internet. The development of professional cooperative

networks goes with the development of the telecommunication and computer networks

and our project is to make these new tools and technologies accessible to the

medical students both during the teaching time in Medical School and during the

training periods at the University Hospital. We have developed a local area

network which communicates between the School of Medicine and the Hospital which

takes advantage of the new Web client-server technology both internally

(Intranet) and externally by access to the National Research Network (RENATER in

France) connected to the Internet network. The address of our public web server

is http:(/)/www.med.univ-rennesl.fr.

PMID: 9476154 [Indexed for MEDLINE]

3873. Int J Med Inform. 1997 Aug;46(1):31-9.

Using the World Wide Web--a new approach to risk identification of diabetes

mellitus.

Baehring TU(1), Schulze H, Bornstein SR, Scherbaum WA.

Author information:

(1)Department of Internal Medicine III, University Hospital of Leipzig, Germany.

baet@server3.medizin.uni-leipzig.de

Diabetes mellitus is a major health problem with a rising tendency world-wide. A

new strategy for risk evaluation and data collection of undiagnosed

non-insulin-dependent diabetes mellitus (NIDDM) using the World Wide Web (WWW) is

presented. An easy-to-handle questionnaire on typical risk factors was converted

into an interactive WWW document. The Internet provides the suitable platform for

the net-based distribution of the questionnaire form as well as the

computer-assisted entering and interpretation of the data

(http:(/)/www.uni-leipzig.de/-diabetes). The analysis program, installed on our

WWW server, interprets the information and sends the assessment on-line back to

the inquiring user PC. The data are also collected anonymously in a database for

epidemiological studies. In the test period, 744 world-wide accesses were

registered: 433 men, mean age 39.8 +/- 14.4 years (range 10-83) and 311 women,

mean age 36.7 +/- 12.7 years (range 11-77). An increased risk for undiagnosed

diabetes was identified to 43.6%. The risk profile of male and female users

showed no significant gender-related differences. Using the WWW technology can

support early detection and adequate treatment of undiagnosed diabetes. This

innovative strategy to screen for a high risk profile is an useful,

cost-effective and up-to-date tool for broad community health education and

epidemiological studies world-wide.

PMID: 9476153 [Indexed for MEDLINE]

3874. J Med Syst. 1997 Aug;21(4):239-48.

A simple WWW interface and quick response system-information query system for

cross-sectional body dimensions.

Hanada E(1), Kenjo Y, Hatae K, Kuromaru R, Antoku Y, Akazawa K, Nose Y.

Author information:

(1)Department of Medical Informatics, Faculty of Medicine, Kyushu University.

We developed a query and analysis system for normal growth measurement of

Japanese children on our WWW server using CGI. It has two subsystems. The first

shows standard height and standard weight calculated by height. This subsystem

can calculate the difference between measured height and the standard along with

deviation and the ratio of measured weight to the standard weight. The second

shows standard height, weight, head circumference, and chest circumference. This

subsystem can calculate differences between the measurements and the standard as

well as deviation. Because of the low amount of output required, very short

turn-around time was required. This system also allows use of the same interface

no matter which brand terminal is used and has wide reusability. This system will

save doctors and nurses the difficulty of looking up a child's data, then having

to make the calculation. We also compare the merits of CGI and Java.

PMID: 9442438 [Indexed for MEDLINE]

3875. Biochim Biophys Acta. 1997 Jul 18;1340(2):253-67.

Computer analysis of phytochrome sequences and reevaluation of the phytochrome

secondary structure by Fourier transform infrared spectroscopy.

Sühnel J(1), Hermann G, Dornberger U, Fritzsche H.

Author information:

(1)Institute of Molecular Biotechnology, Jena, Germany.

A repertoire of various methods of computer sequence analysis was applied to

phytochromes in order to gain new insights into their structure and function. A

statistical analysis of 23 complete phytochrome sequences revealed regions of

non-random amino acid composition, which are supposed to be of particular

structural or functional importance. All phytochromes other than phyD and phyE

from Arabidopsis have at least one such region at the N-terminus between residues

2 and 35. A sequence similarity search of current databases indicated striking

homologies between all phytochromes and a hypothetical 84.2-kDa protein from the

cyanobacterium Synechocystis. Furthermore, scanning the phytochrome sequences for

the occurrence of patterns defined in the PROSITE database detected the signature

of the WD repeats of the beta-transducin family within the functionally important

623-779 region (sequence numbering of phyA from Avena) in a number of

phytochromes. A multiple sequence alignment performed with 23 complete

phytochrome sequences is made available via the IMB Jena World-Wide Web server

(http://www.imb-jena.de/PHYTO.html). It can be used as a working tool for future

theoretical and experimental studies. Based on the multiple alignment striking

sequence differences between phytochromes A and B were detected directly at the

N-terminal end, where all phytochromes B have an additional stretch of 15-42

amino acids. There is also a variety of positions with totally conserved but

different amino acids in phytochromes A and B. Most of these changes are found in

the sequence segment 150-200. It is, therefore, suggested that this region might

be of importance in determining the photosensory specificity of the two

phytochromes. The secondary structure prediction based on the multiple alignment

resulted in a small but significant beta-sheet content. This finding is confirmed

by a reevaluation of the secondary structure using FTIR spectroscopy.

PMID: 9252112 [Indexed for MEDLINE]

3876. Mamm Genome. 1997 Jul;8(7):467-71.

HOSEpipe--a WWW-hosted data management and analysis system for STS content

mapping projects.

Strivens MA(1), Middlehurst P, Brown SD, Denny P.

Author information:

(1)MRC Mouse Genome Centre and MRC Mammalian Genetics Unit, Harwell, OX11 ORD,

UK.

We have developed a data management system, 'HOSEpipe' (High Output STS

Evaluation pipeline) to aid sample tracking and data analysis in STS content

mapping projects. The system is based around a World Wide Web (WWW) server that

provides a number of pages including forms for sample processing and data entry

accessible via a standard WWW browser application. The system is split into two

main modules: firstly, a sequence evaluation and annotation module that takes de

novo sequence for a potential STS, screens it against existing STSs and DNA

sequence databases, followed by appropriate primer sequence design; secondly, a

module that handles YAC library STS screening and includes facilities for both

sample tracking and experimental data analysis. We present the design and

rationale of the HOSEpipe system and its development to support a whole

chromosomal physical mapping project. This software and design approach is

potentially applicable to physical mapping projects of varying sizes and

resolution and to similar projects, such as sample sequencing and the

construction of sequence-ready maps.

PMID: 9195989 [Indexed for MEDLINE]

3877. Nihon Koshu Eisei Zasshi. 1997 Jul;44(7):518-22.

[Application of Internet technology in public health].

[Article in Japanese]

Satoh T(1), Takahashi K, Yahata K, Nakagawa S, Wojtczak A, Takizawa Y, Tajima N,

Kohyama A, Akazawa S, Higashi T, Yamaguchi N, Sekikawa A.

Author information:

(1)Department of Hygiene and Public Health, Tokyo Women's Medical College.

Recent advances in telecommunication technology have been enormous. Application

of this technology in public health has the potential to markedly improve global

health through better surveillance and information systems. With this assumption

the GHNet was established in 1994 by representatives from academia, WHO, Pan

American Health Organization, the World Bank, NASA, IBM, and AT & T. The GHNet

consists of seven components: 1) promotion of networking with the Internet among

people in public health; 2) disease tele-monitoring; 3) distance learning system

with the internet; 4) connection of non-governmental health organizations; 5)

training cyberdocs who are educated in both public health and telecommunications;

6) establishment of an electronic scientific research server; and 7) a home page

on the World Wide Web (WWW). In order to effectively incorporate the Internet

into the field, connectivity and knowing how to use it are of critical concern.

More and more facilities are connected to the Internet in Japan. However, few

courses teaching how to utilize the Internet are provided for people in this

field. An Internet training course for people in public health was held as joint

venture of the World Health Organization (WHO) and the Global Health Network

(GHNet) on October 31, 1996, at the 55th Annual Meeting of Japanese Society of

Public Health. Most of the participants for the course were from local public

health departments and very few had previous experience with the Internet before

the course. During this course participants learned how to use e-mail, how to

find health resources on the WWW, how to construct a home page, and how the

Internet could be utilized to improve public health, with their computers

actually hooked to the Internet. From this experience, we found that this kind of

course is feasible and beneficial and hope that this course would serve as a

model for training people in public health.

PMID: 9314706 [Indexed for MEDLINE]

3878. Appl Environ Microbiol. 1997 Jun;63(6):2338-46.

A new computational method for detection of chimeric 16S rRNA artifacts generated

by PCR amplification from mixed bacterial populations.

Komatsoulis GA(1), Waterman MS.

Author information:

(1)Department of Mathematics, University of Southern California, Los Angeles

90089-1113, USA. gkoma@hto.usc.edu

A new computational method (chimeric alignment) has been developed to detect

chimeric 16S rRNA artifacts generated during PCR amplification from mixed

bacterial populations. In contrast to other nearest-neighbor methods (e.g.,

CHECK\_CHIMERA) that define sequence similarity by k-tuple matching, the chimeric

alignment method uses the score from dynamic programming alignments. Further, the

chimeric alignments are displayed to the user to assist in sequence

classification. The distribution of improvement scores for 500 authentic,

nonchimeric sequences and 300 artificial chimeras (constructed from authentic

sequences) was used to study the sensitivity and accuracy of both chimeric

alignment and CHECK\_CHIMERA. At a constant rate of authentic sequence

misclassification (5%), chimeric alignment incorrectly classified 13% of the

artificial chimeras versus 14% for CHECK\_CHIMERA. Interestingly, only 1% of

nonchimeras and 10% of chimeras were misclassified by both programs, suggesting

that optimum performance is obtained by using the two methods to assign sequences

to three classes: high-probability nonchimeras, high-probability chimeras, and

sequences that need further study by other means. This study suggests that

k-tuple-based matching methods are more sensitive than alignment-based methods

when there is significant parental sequence similarity, while the opposite

becomes true as the sequences become more distantly related. The software and a

World Wide Web-based server are available at

http://www-hto.usc.edu/software/mglobal CHI.

PMCID: PMC168526

PMID: 9172353 [Indexed for MEDLINE]

3879. Comput Appl Biosci. 1997 Jun;13(3):249-56.

Match-Box\_server: a multiple sequence alignment tool placing emphasis on

reliability.

Depiereux E(1), Baudoux G, Briffeuil P, Reginster I, De Bolle X, Vinals C,

Feytmans E.

Author information:

(1)Department of Biology, Facultés Universitaires Notre-Dame de la Paix, Namur,

Belgium. eric.depiereux@fundp.ac.be

MOTIVATION: The Match-Box software comprises protein sequence alignment tools

based on strict statistical thresholds of similarity between protein segments.

The method circumvents the gap penalty requirement: gaps being the result of the

alignment and not a governing parameter of the procedure. The reliable conserved

regions outlined by Match-Box are particularly relevant for homology modelling of

protein structures, prediction of essential residues for site-directed

mutagenesis and oligonucleotide design for cloning homologous genes by polymerase

chain reaction (PCR).

RESULTS: The method produces reliable results, as assessed by tests performed on

protein families of known structures and of low sequence similarity. A

reliability score is computed in relation to a threshold of similarity

progressively raised to extend the aligned regions to their maximal length, up to

the significance limit of matching segments. The score obtained at each position

is printed below the sequences and allows a discriminant reading of each aligned

region.

AVAILABILITY: Sequences may be submitted to a Web server at

http://www.fundp.ac.be/sciences/biologie/bms/+ ++matchbox\_submit.html or sent by

e-mail to matchbox/biq.fundp.ac.be (help available by just mailing help).

PMID: 9183529 [Indexed for MEDLINE]

3880. Radiol Med. 1997 Jun;93(6):743-50.

[Server World-Wide Web on the Internet for the provision of clinical cases and

digital radiologic images for training and continuing education in radiology].

[Article in Italian]

Sparacia G(1), Tartamella M, Finazzo M, Bartolotta T, Brancatelli G, Banco A, Lo

Casto A, La Tona G, Bentivegna E.

Author information:

(1)Istituto di Radiologia P. Cignolini, Università di Palermo.

sparacia@cartesio.cere.pa.cnr.it

The Internet, as a global computer network, provides opportunities to make

available multimedia educational materials, such as teaching files and image

databases, that can be accessed using "World-Wide Web" client browser to provide

continuing medical education. Since August, 1995, at the Institute of

Radiology-University of Palermo, we developed a World-Wide Web server on the

Internet to provide a collection of interactive radiology educational resources

such as teaching files and image database for continuing medical education in

radiology. Our server is based on a UNIX workstation connected to the Internet

via our campus Ethernet network and reachable at the uniform resource locator

(URL) address: http:/(/)mbox.unipa.it/approximately radpa/ radpa.html. Digital CT

and MR images for teaching files and image database are downloaded through an

Ethernet local area network from a GE Advantage Windows workstation. US images

will be acquired on-line through a video digitizing board. Radiographs will be

digitized by means of a Charge Coupled Device (CCD) scanner. To set up teaching

files, image database and all other documents, we use the standard "HyperText

Markup Language" (HTML) to edit the documents, and the Graphics Interchange

Format (GIF) or Joint Photographic Expert Group (JPEG) format to store the

images. Nine teaching files are presently available on the server, together with

49 images in the database, a list of international radiological servers, a

section devoted to the museum of radiology hosted by our Institute, the

electronic version of the Journal Eido Electa. In the first 12 months of public

access through the Internet, 12,280 users accessed the server worldwide: 45% of

them to retrieve teaching files; 35% to retrieve images from the database; the

remaining 20% to retrieve other documents. Placing teaching files and image

database on a World-Wide Web server makes these cases more available to residents

and radiologists to provide continuing medical education in radiology.

PMID: 9411524 [Indexed for MEDLINE]

3881. J Chem Inf Comput Sci. 1997 May-Jun;37(3):417-24.

The PRINTS database of protein fingerprints: a novel information resource for

computational molecular biology.

Attwood TK(1), Avison H, Beck ME, Bewley M, Bleasby AJ, Brewster F, Cooper P,

Degtyarenko K, Geddes AJ, Flower DR, Kelly MP, Lott S, Measures KM, Parry-Smith

DJ, Perkins DN, Scordis P, Scott D, Worledge C.

Author information:

(1)Department of Biochemistry and Molecular Biology, University College London,

UK.

PRINTS is a compendium of protein motif fingerprints derived from the OWL

composite sequence database. Fingerprints are groups of motifs within sequence

alignments whose conserved nature allows them to be used as signatures of family

membership. Fingerprints inherently offer improved diagnostic reliability over

single motif methods by virtue of the mutual context provided by motif neighbors.

To date, 650 fingerprints have been constructed and stored in PRINTS, the size of

which has doubled in the last 2 years. The current version, 14.0, encodes 3500

motifs, covering a range of globular and membrane proteins, modular polypeptides,

and so on. The database is now accessible via the UCL Bioinformatics Server on

http:@ www.biochem.ucl.ac.uk/bsm/dbbrowser/. We describe here progress with the

database, its compilation and interrogation software, and its Web interface.

PMID: 9177000 [Indexed for MEDLINE]

3882. FEBS Lett. 1997 Apr 7;406(1-2):69-74.

Distribution of sequence-dependent curvature in genomic DNA sequences.

Gabrielian A(1), Vlahovicek K, Pongor S.

Author information:

(1)International Centre for Genetic Engineering and Biotechnology (ICGEB), Area

Science Park, Trieste, Italy.

The distribution of inherent, sequence-dependent curvature was calculated for a

number of prokaryotic (M. genitalium, H. influenzae, M. jannaschii), viral

(adenovirus 2, equine herpes virus 1), phage (M13, lambda), eukaryotic (S.

cerevisiae) and mitochondrial genomes as well as E. coli and human genomic

fragments. The genomic averages are in the range of 6-8 degrees/helical turn and

only about 20% of DNA is curved less than 3 degrees/helical turn. The prokaryotes

and phages appear to have a consistently higher frequency of curved DNA in their

genomes than the other genomes tested. Long, highly curved segments, similar to

artificially designed curved DNA, are apparently absent from the genomes. Short,

curved segments, differing in G+C content may provide environmentally modulated

conformational signals for gene regulation. A WWW-server was constructed for the

prediction of curved sites from DNA sequences

(http://icgeb.trieste.it/dna/curve\_it.html/)..

PMID: 9109388 [Indexed for MEDLINE]

3883. Electrophoresis. 1997 Mar-Apr;18(3-4):599-604.

Renal cell carcinoma and normal kidney protein expression.

Sarto C(1), Marocchi A, Sanchez JC, Giannone D, Frutiger S, Golaz O, Wilkins MR,

Doro G, Cappellano F, Hughes G, Hochstrasser DF, Mocarelli P.

Author information:

(1)University Department of Clinical Pathology, Desio Hospital, Milan, Italy.

sarto@desiolab.unimi.it

Renal cell carcinoma (RCC), a human kidney cancer from the proximal tubular

epithelium, accounts for about 3% of adult malignancies. Molecular and

cytogenetic analysis have highlighted deletions, translocations, or loss of

heterozygosity in the 3p21-p26, a putative RCC locus, as well as in 6q, 8p, 9pq,

and 14pq. Studies on phenotypic expression of human kidney tissue and on

post-translational modifications in RCC have not yet provided a marker for early

renal cell carcinoma diagnosis. Current diagnostic methods do not help to detect

the tumor before advanced stages. We therefore used two-dimensional

polyacrylamide gel electrophoresis (2-D PAGE) to study normal and tumor kidney

tissues in ten patients suffering from RCC. A human kidney protein map in the

SWISS-2DPAGE database accessible through the ExPASy WWW Molecular Biology Server

was established. Of 2789 separated polypeptides, 43 were identified by gel

comparison, amino acid analysis, N-terminal sequencing, and/or immunodetection.

The comparison between normal and tumor kidney tissues showed four polypeptides

to be absent in RCC. One of them was identified as ubiquinol cytochrome c

reductase (UQCR), whose locus has elsewhere been tentatively assigned to

chromosome 19p12 or chromosome 22. A second polypeptide was identified as

mitochondrial NADH-ubiquinone oxido-reductase complex I whose locus is located on

chromosome 18p11.2 and chromosome 19q13.3. These result suggest that the lack of

UQCR and of mitochondrial NADH-ubiquinone oxidoreductase complex I expression in

RCC may be caused by unknown deletions, or by changes in gene transcription or

translation. It might indicate that mitochondrial disfunction plays a major role

in RCC genesis or evolution.

DOI: 10.1002/elps.1150180343

PMID: 9150947 [Indexed for MEDLINE]

3884. Electrophoresis. 1997 Mar-Apr;18(3-4):498-501.

Large-scale protein modelling and integration with the SWISS-PROT and

SWISS-2DPAGE databases: the example of Escherichia coli.

Peitsch MC(1), Wilkins MR, Tonella L, Sanchez JC, Appel RD, Hochstrasser DF.

Author information:

(1)Geneva Biomedical Research Institute, Glaxo Wellcome Research and Development,

Plan-les-Ouates/Genève, Switzerland. mcp13936@ggr.co.uk

Knowledge-based molecular modelling of proteins has proven useful in many

instances, including the rational design of mutagenesis experiments, but it has

generally been limited by the availability of expensive computer hardware and

software. To overcome these limitations, we developed the SWISS-MODEL server for

automated knowledge-based protein modelling. The SWISS-MODEL server uses the

Brookhaven Protein Data Bank as a source of structural information and

automatically generates protein models for sequences which share significant

similarities with at least one protein of known three-dimensional structure. We

have now used the software framework of the server to generate large collections

of protein models, and established the SWISS-MODEL Repository, a new database for

automatically generated and theoretical protein models. This repository is

directly integrated with the SWISS-PROT and SWISS-2DPAGE databases through the

ExPASy World Wide Web server (URL is http://expasy.hcuge.ch). Here we present an

illustration of this process by an application to the Escherichia coli sequences.

DOI: 10.1002/elps.1150180326

PMID: 9150930 [Indexed for MEDLINE]

3885. Electrophoresis. 1997 Mar-Apr;18(3-4):484-90.

An on-line two-dimensional polyacrylamide gel electrophoresis protein database of

adult Drosophila melanogaster.

Ericsson C(1), Pethö Z, Mehlin H.

Author information:

(1)Karolinska Institutet, Department of Cell and Molecular Biology, Medical Nobel

Institute, Stockholm, Sweden. ericsson@acsu.buffalo.edu

An annotated two-dimensional polyacrylamide gel electrophoresis (2-D PAGE)

protein database of adult Drosophila melanogaster has been constructed, based on

the protein patterns of heads, thoraces and abdomens of adult male and female

Drosophila melanogaster. About 1200 major protein spots are catalogued. Common

proteins, found in all body parts, as well as bodypart- and sex-specifically

expressed proteins are reported. Of the major proteins, 91, or 7.5%, are

differentially expressed in the two sexes or in different body parts, at least in

part reflecting specific functional requirements. At the present time 43

proteins, or about 3.5% of the detected proteins, have been identified. These

data can be accessed interactively from our World Wide Web (WWW) server through

clickable inline gel images and hypertext links. Identified protein spots are

cross-referenced, through hypertext links, to the SWISS-PROT annotated database

of protein primary sequences and the Fly-Base database of Drosophila genomic

data. Our reference gels can be used to gain immediate access to protein spot

identify and to the pattern of differentially expressed proteins in Drosophila

melanogaster. The work presented in this article ties together information from

protein 2-D PAGE, molecular biology and genetics and offers a uniform way to

access this large volume of data.

DOI: 10.1002/elps.1150180324

PMID: 9150928 [Indexed for MEDLINE]

3886. Electrophoresis. 1997 Mar-Apr;18(3-4):403-8.

Detailed peptide characterization using PEPTIDEMASS--a World-Wide-Web-accessible

tool.

Wilkins MR(1), Lindskog I, Gasteiger E, Bairoch A, Sanchez JC, Hochstrasser DF,

Appel RD.

Author information:

(1)Central Clinical Chemistry Laboratory, Geneva University Hospital,

Switzerland. marc.wilkins@dim.hcuge.ch

In peptide mass fingerprinting, there are frequently peptides whose masses cannot

be explained. These are usually attributed to either a missed cleavage site

during the chemical or enzymatic cutting process, the lack of reduction and

alkylation of a protein, protein modifications like the oxidation of methionine,

or the presence of protein post-translational modifications. However, they could

equally be due to database errors, unusual splicing events, variants of a protein

in a population, or artifactual protein modifications. Unfortunately the

verification of each of these possibilities can be tedious and time-consuming. To

better utilize annotated protein databases for the understanding of peptide mass

fingerprinting data, we have written the program "PEPTIDEMASS". This program

generates the theoretical peptide masses of any protein in the SWISS-PROT

database, or of any sequence specified by the user. If the sequence is derived

from the SWISS-PROT database, the program takes into account any annotations for

that protein in order to generate the peptide masses. In this manner, the user

can obtain the predicted masses of peptides from proteins which are known to have

signal sequences, propeptides, transit peptides, simple post-translational

modifications, and disulfide bonds. Users are also warned if any peptide masses

are subject to change from protein isoforms, database conflicts, or an mRNA

splicing variation. The program is freely accessible to the scientific community

via the ExPASy World Wide Web server, at the URL address:

http://www.expasy.ch/www/tools.html.

DOI: 10.1002/elps.1150180314

PMID: 9150918 [Indexed for MEDLINE]

3887. Methods Inf Med. 1997 Feb;36(2):149-53.

Computer-based training and electronic publishing in the health sector: tools and

trends.

Schulz S(1), Schrader U, Klar R.

Author information:

(1)Department of Medical Informatics, University of Freiburg, Germany.

schulz@amil.ukl.uni-freiburg.de

CBT (computer-based training) applications and hypermedia publications are two

different approaches to the utilisation of computers in medical education.

Medical CBT software continues to play a minor role in spite of the increasing

availability, whereas hypermedia have become very popular through the World Wide

Web (WWW). Based on the HTML format they can be designed by non-programmers using

inexpensive tools while the production of CBT applications requires programming

expertise. HTML documents can be easily developed to be distributed by a

web-server or to run as local applications. In developed countries CBT and

hypermedia have to compete with an abundance of printed or audio-visual media and

a wealth of lectures, conferences, etc., whereas in developing countries these

media are scarce and expensive. Here CBT programs, and hypermedia publications in

particular, may be a cost-effective way to improve quality of education in the

health sector.

PMID: 9242015 [Indexed for MEDLINE]

3888. Hum Brain Mapp. 1997;5(4):238-42. doi:

10.1002/(SICI)1097-0193(1997)5:4<238::AID-HBM6>3.0.CO;2-4.

Automated labeling of the human brain: a preliminary report on the development

and evaluation of a forward-transform method.

Lancaster JL(1), Rainey LH, Summerlin JL, Freitas CS, Fox PT, Evans AC, Toga AW,

Mazziotta JC.

Author information:

(1)Research Imaging Center, University of Texas Health Science Center at San

Antonio, San Antonio, Texas 78284-6240, USA.

A forward-transform method for retrieving brain labels from the 1988 Talairach

Atlas using x-y-z coordinates is presented. A hierarchical volume-occupancy

labeling scheme was created to simplify the organization of atlas labels using

volume and subvolumetric components. Segmentation rules were developed to define

boundaries that were not given explicitly in the atlas. The labeling scheme and

segmentation rules guided the segmentation and labeling of 160 contiguous regions

within the atlas. A unique three-dimensional (3-D) database label server called

the Talairach Daemon (http://ric.uthscsa.edu/projects) was developed for serving

labels keyed to the Talairach coordinate system. Given an x-y-z Talairach

coordinate, a corresponding hierarchical listing of labels is returned by the

server. The accuracy and precision of the forward-transform labeling method is

now under evaluation.

Copyright (c) 1997 Wiley-Liss, Inc.

DOI: 10.1002/(SICI)1097-0193(1997)5:4<238::AID-HBM6>3.0.CO;2-4

PMCID: PMC2860189

PMID: 20408222

3889. Nucleic Acids Res. 1997 Jan 1;25(1):240-3.

The SBASE protein domain library, release 5.0: a collection of annotated protein

sequence segments.

Fábián P(1), Murvai J, Hátsági Z, Vlahovicek K, Hegyi H, Pongor S.

Author information:

(1)ABC Institute for Biochemistry and Protein Research, 2100 Gödöllö, Hungary.

SBASE 5.0 is the fifth release of SBASE, a collection of annotated protein domain

sequences that represent various structural, functional, ligand-binding and

topogenic segments of proteins. SBASE was designed to facilitate the detection of

functional homologies and can be searched with standard database-search programs.

The present release contains over 79863 entries provided with standardized names

and is cross-referenced to all major sequence databases and sequence pattern

collections. The information is assigned to individual domains rather than to

entire protein sequences, thus SBASE contains substantially more cross-references

and links than do the protein sequence databases. The entries are clustered into

>16 000 groups in order to facilitate the detection of distant similarities.

SBASE 5.0 is freely available by anonymous 'ftp' file transfer from

<ftp.icgeb.trieste.it >. Automated searching of SBASE with BLAST can be carried

out with the WWW-server <http://www.icgeb.trieste.it/sbase/ >. and with the

electronic mail server <sbase@icgeb.trieste.it >which now also provides a graphic

representation of the homologies. A related WWW-server

<http://www.abc.hu/blast.html > and e-mail server <domain@hubi.abc.hu > predicts

SBASE domain homologies on the basis of SWISS-PROT searches.

PMCID: PMC146372

PMID: 9016545 [Indexed for MEDLINE]

3890. Nucleic Acids Res. 1997 Jan 1;25(1):212-7.

Novel developments with the PRINTS protein fingerprint database.

Attwood TK(1), Beck ME, Bleasby AJ, Degtyarenko K, Michie AD, Parry-Smith DJ.

Author information:

(1)Department of Biochemistry and Molecular Biology, University College London,

Gower Street, London WCIE 6BT, UK. attwood@bsm.bioc.ucl.ac.uk

The PRINTS database of protein family 'fingerprints' is a diagnostic resource

that complements the PROSITE dictionary of sites and patterns. Unlike regular

expressions, fingerprints exploit groups of conserved motifs within sequence

alignments to build characteristic signatures of family membership. Thus

fingerprints inherently offer improved diagnostic reliability by virtue of the

mutual context provided by motif neighbours. To date, 600 fingerprints have been

constructed and stored in PRINTS, representing a 50% increase in the size of the

database in the last year. The current version, 13.0, encodes approximately 3000

motifs, covering a range of globular and membrane proteins, modular polypeptides,

and so on. The database is accessible via UCL's Bioinformatics World Wide Web

(WWW) server at http://www.biochem.ucl.ac.uk/bsm/dbbrowser / . We describe here

progress with the database, its Web interface, and a recent exciting development:

the integration of a novel colour alignment editor

(http://www.biochem.ucl.ac.uk/bsm/dbbrowser++ +/CINEMA ), which allows

visualisation and interactive manipulation of PRINTS alignments over the

Internet.

PMCID: PMC146411

PMID: 9016538 [Indexed for MEDLINE]

3891. Nucleic Acids Res. 1997 Jan 1;25(1):117-22.

Database on the structure of large ribosomal subunit RNA.

De Rijk P(1), Van de Peer Y, De Wachter R.

Author information:

(1)Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1,

B-2610 Antwerpen, Belgium.

The latest release of the large ribosomal subunit RNA database contains 429

sequences. All these sequences are aligned, and incorporate secondary structure

information. The rRNA WWW Server at URL http://rrna.uia.ac.be/ provides

researchers with an easily accessible resource to obtain the data in this

database in a number of computer-readable formats. A new query interface has been

added to the server. If necessary, the data can also be obtained by anonymous ftp

from the same site.

PMCID: PMC146373

PMID: 9016517 [Indexed for MEDLINE]

3892. Nucleic Acids Res. 1997 Jan 1;25(1):109-11.

The RDP (Ribosomal Database Project).

Maidak BL(1), Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, Woese CR.

Author information:

(1)Department of Microbiology, University of Illinois, B103 C&LSL, 601 South

Goodwin Avenue, Urbana, IL 61801, USA. b-maidak@uiuc.edu

The Ribosomal Database Project (RDP) is a curated database that offers

ribosome-related data, analysis services and associated computer programs. The

offerings include phylogenetically ordered alignments of ribosomal RNA (rRNA)

sequences, derived phylogenetic trees, rRNA secondary structure diagrams, and

various software for handling, analyzing and displaying alignments and trees. The

data are available via anonymous FTP (rdp.life.uiuc.edu), electronic mail

(server@rdp.life.uiuc.edu), gopher (rdpgopher.life.uiuc.edu) and WWW

(http://rdpwww.life.uiuc.edu/ ). The electronic mail and WWW servers provide

ribosomal probe checking, approximate phylogenetic placement of user-submitted

sequences, screening for possible chimeric rRNA sequences, automated alignment,

and a suggested placement of an unknown sequence on an existing phylogenetic

tree.

PMCID: PMC146422

PMID: 9016515 [Indexed for MEDLINE]

3893. Nucleic Acids Res. 1997 Jan 1;25(1):28-30.

MIPS: a database for protein sequences, homology data and yeast genome

information.

Mewes HW(1), Albermann K, Heumann K, Liebl S, Pfeiffer F.

Author information:

(1)MIPS, Max-Planck-Institut für Biochemie, Am Klopferspitz 18, D-82152

Martinsried, Germany. mewes@mips.biochem.mpg.de

The MIPS group (Martinsried Institute for Protein Sequences) at the

Max-Planck-Institute for Biochemistry, Martinsried near Munich, Germany,

collects, processes and distributes protein sequence data within the framework of

the tripartite association of the PIR-International Protein Sequence Database

(,). MIPS contributes nearly 50% of the data input to the PIR-International

Protein Sequence Database. The database is distributed on CD-ROM together with

PATCHX, an exhaustive supplement of unique, unverified protein sequences from

external sources compiled by MIPS. Through its WWW server

(http://www.mips.biochem.mpg.de/ ) MIPS permits internet access to sequence

databases, homology data and to yeast genome information. (i) Sequence similarity

results from the FASTA program () are stored in the FASTA database for all

proteins from PIR-International and PATCHX. The database is dynamically

maintained and permits instant access to FASTA results. (ii) Starting with FASTA

database queries, proteins have been classified into families and superfamilies

(PROT-FAM). (iii) The HPT (hashed position tree) data structure () developed at

MIPS is a new approach for rapid sequence and pattern searching. (iv) MIPS

provides access to the sequence and annotation of the complete yeast genome (),

the functional classification of yeast genes (FunCat) and its graphical display,

the 'Genome Browser' (). A CD-ROM based on the JAVA programming language

providing dynamic interactive access to the yeast genome and the related protein

sequences has been compiled and is available on request.

PMCID: PMC146421

PMID: 9016498 [Indexed for MEDLINE]

3894. Proc AMIA Annu Fall Symp. 1997:500-4.

Standardized problem list generation, utilizing the Mayo canonical vocabulary

embedded within the Unified Medical Language System.

Elkin PL(1), Mohr DN, Tuttle MS, Cole WG, Atkin GE, Keck K, Fisk TB, Kaihoi BH,

Lee KE, Higgins MC, Suermondt HJ, Olson N, Claus PL, Carpenter PC, Chute CG.

Author information:

(1)Department of Area Internal Medicine, Mayo Clinic, Rochester, MN, USA.

VOCABULARY: The Mayo problem list vocabulary is a clinically derived lexicon

created from the entries made to the Mayo Clinic's Master Sheet Index and the

problem list entries made to the Impression/ Report/Plan section of the Clinical

Notes System over the last three years. The vocabulary was reduced by eliminating

repetition including lexical variants, spelling errors, and qualifiers

(Administrative or Operational terms). Qualifiers are re-coordinated with other

terms, at run-time, which greatly increased the number of input strings which our

system is capable of recognizing.IMPLEMENTATION: The Problem Manager is

implemented using standard windows tools in a Windows NT environment. The

interface is designed using Object Pascal. HTTP calls are passed over the World

Wide Web to a UNIX based vocabulary server. The server returns a document, which

is read into Object Pascal structures, parsed, filtered and displayed.

STUDY: This paper reports the results of a recent Usability Trial focused on

assessing the viability of this mechanism for standardized problem entry. Eight

clinicians engaged in eleven scenarios and responded as to their satisfaction

with the systems performance. These responses were observed, videotaped and

tabulated. Clinicians in this study were able to find acceptable diagnoses in

91.1% of the scenarios. The response time was acceptable in 92.5% of the

scenarios. The presentation of related terms was stated to be useful in at least

one scenario by seven of the eight participants. All clinicians wanted to make

use of shortcuts which would minimize the amount of typing necessary to encode

the concept they were searching for (e.g. Abbreviations, Word Completion).

CONCLUSIONS: Clinicians are willing to choose a canonical term from a suggested

list (as opposed to their own wording). Clinicians want an "intelligent" system,

which would suggest terms within a category (e.g. Types of "Migraine"). They are

able to make functional use of our system, in its current state of development.

Finally, all clinicians appreciate the value of encoding their problems in a

standardized vocabulary, toward improved research, education and practice.

PMCID: PMC2233586

PMID: 9357676 [Indexed for MEDLINE]

3895. Proc Int Conf Intell Syst Mol Biol. 1997;5:234-6.

Large scale protein modelling and model repository.

Peitsch MC(1).

Author information:

(1)Geneva Biomedical Research Institute, Plan-les-Ouates/Geneva, Switzerland.

mcp13936@ggr.co.uk

Knowledge-based molecular modelling of proteins has proven useful in many

instances including the rational design of mutagenesis experiments, but it has

been generally limited by the availability of expensive computer hardware and

software. To overcome these limitations, we have developed the SWISS-MODEL server

for automated knowledge-based protein modelling. The SWISS-MODEL server uses the

Brookhaven Protein Data Bank as a source of structural information and

automatically generates protein models for sequences which share significant

similarities with at least one protein of known 3D-structure. We now use the

software framework of the server to generate large collections of protein models.

To store these models, we have established the SWISS-MODEL Repository, a new

database for protein models generated by theoretical approaches. This repository

is directly integrated with SWISS-PROT and other databases through the ExPASy

World-Wide Web server (URL is http:(/)/www.expasy.ch).

PMID: 9322042 [Indexed for MEDLINE]

3896. Proc Int Conf Intell Syst Mol Biol. 1997;5:214-7.

PDB-REPRDB: a database of representative protein chains in PDB (Protein Data

Bank).

Noguchi T(1), Onizuka K, Akiyama Y, Saito M.

Author information:

(1)Parallel Application Laboratory, Tsukuba Research Center, Real World Computing

Partnership, Japan. noguchi@trc.rwcp.or.jp

We propose a novel set of 'representative' protein chains in PDB, where not only

sequential but also structural similarities are taken into account. Hobohm et al.

have already proposed "PDB\_SELECT", which eliminates redundant chains based

solely on sequence similarity. "PDB\_SELECT" is frequently updated and the latest

version is available at EMBL WWW server. In our set of entries "PDB-REPRDB,"

however, structural similarities are also considered, in order not to overlook

local conformation diversity within a group of sequentially similar chains. Our

set guarantees that every representative is the best among that similar protein

group, regarding experimental or structure-determination quality (i.e. resolution

and R-value). The first version (based on PDB Release 70) of PDB-REPRDB was

released in 1995 and the second version (PDB Release 78) will be available by

April 1997.

PMID: 9322039 [Indexed for MEDLINE]

3897. Proc Int Conf Intell Syst Mol Biol. 1997;5:187-90.

ANOLEA: a www server to assess protein structures.

Melo F(1), Devos D, Depiereux E, Feytmans E.

Author information:

(1)Department of Biology, Facultés Universitaires Notre Dame de la Paix,

Bruxelles, Belgium.

ANOLEA (Atomic Non-Local Environment Assessment) is a www server that performs

energy calculations at the atomic level in protein structures. The calculations

involve the non-local interactions between all the heavy atoms of the twenty

standard amino acids in the molecule. The input of the server is a PDB file

containing one or more protein chains. The output is an energy profile, which

gives an energy value for each amino acid of the protein. High energy zones

(HEZs) in the profile correlate with errors or with potential interacting zones

of proteins. The output of the server also displays the structure in three

dimensions, pointing out the high energy amino acids in the protein. This option

requires the CHIME plug-in, which is freely available on Internet and makes

possible, in real time, to rotate, translate and change the point of view and

presentation of the molecule in three dimensions. Thus, a fast analysis of a

protein structure can be done using a personal computer connected to Internet.

The server is available at: http:@www.fundp.ac.be/pub/ANOLEA.html.

PMID: 9322034 [Indexed for MEDLINE]

3898. Protein Eng. 1997 Jan;10(1):1-6.

Identification of prokaryotic and eukaryotic signal peptides and prediction of

their cleavage sites.

Nielsen H(1), Engelbrecht J, Brunak S, von Heijne G.

Author information:

(1)Department of Chemistry, Technical University of Denmark, Lyngby, Denmark.

We have developed a new method for the identification of signal peptides and

their cleavage sites based on neural networks trained on separate sets of

prokaryotic and eukaryotic sequence. The method performs significantly better

than previous prediction schemes and can easily be applied on genome-wide data

sets. Discrimination between cleaved signal peptides and uncleaved N-terminal

signal-anchor sequences is also possible, though with lower precision.

Predictions can be made on a publicly available WWW server.

PMID: 9051728 [Indexed for MEDLINE]

3899. Comput Appl Biosci. 1996 Dec;12(6):519-24.

Correspondence discriminant analysis: a multivariate method for comparing classes

of protein and nucleic acid sequences.

Perrière G(1), Lobry JR, Thioulouse J.

Author information:

(1)Laboratoire de Biométrie, Génétique et Biologie des Populations, UMR CNRS n

5558, Université Claude Bernard, Lyon, Villeurbanne, France.

perriere.lobry.thioulou@biomserv.univ-lyonl.fr

This report describes two applications of a multivariate method for studying

classes of nucleotide or protein sequences: correspondence discriminant analysis

(CDA). The first example is the discrimination between Escherichia coli proteins

according to their subcellular location (membrane, cytoplasm and periplasm). The

high resolution of the method made it possible to predict the subcellular

location of E.coli proteins for whom this information is not known. The second

example is discrimination between the coding sequences of leading and lagging

strands in four bacteria: Mycoplasma genitalium, Haemophilus influenzae, E.coli

and Bacillus subtilis. The programs used for computing the analysis are

integrated in a publicly available package that runs on MacOS 7.x or Windows 95

operating systems (http:/(/)biomserv.univ-lyonl.fr/ADE-4.html). These programs

are also accessible through our World Wide Web server

(http:/(/)biomserv.univ-lyonl.fr/Net Mul.html).

PMID: 9021271 [Indexed for MEDLINE]

3900. Comput Appl Biosci. 1996 Dec;12(6):507-10.

LALNVIEW: a graphical viewer for pairwise sequence alignments.

Duret L(1), Gasteiger E, Perrière G.

Author information:

(1)Department of Medical Biochemistry, University of Geneva, Switzerland.

duret@dim.hcuge.ch

LALNVIEW is a graphical program for visualising local alignments between two

sequences (protein or nucleic acids). Sequences are represented by coloured

rectangles to give an overall picture of their similarities. LALNVIEW can display

sequence features (exon, intron, active site, domain, propeptide, etc.) along

with the alignment. When using LALNVIEW through our Web servers, sequence

features are automatically extracted from database annotations (SWISS-PROT,

GenBank, EMBL or HOVERGEN) and displayed with the alignment. LALNVIEW is a useful

tool for analysing pairwise sequence alignments and for making the link between

sequence homology and what is known about the structure or function of sequences.

LALNVIEW executables for UNIX, Macintosh and PC computers are freely available

from our server (http:// expasy.hcuge.ch/sprot/lalnview.html).

PMID: 9021269 [Indexed for MEDLINE]

3901. Control Clin Trials. 1996 Dec;17(6):476-93.

A World Wide Web-based user interface for a data management system for use in

multi-institutional clinical trials--development and experimental operation of an

automated patient registration and random allocation system.

Kiuchi T(1), Ohashi Y, Konishi M, Bandai Y, Kosuge T, Kakizoe T.

Author information:

(1)Department of Epidemiology, Faculty of Medicine, University of Tokyo, Japan.

We have employed the Hypertext Transfer Protocol (HTTP) and Hypertext Markup

Language (HTML) to develop an automated patient registration and random

allocation system for use in a multi-institutional clinical trial. We made it

available on-line to World Wide Web clients in each hospital through a user

friendly graphical user interface. During experimental operation, the physicians

found it satisfactory from the viewpoint of both ease of operation and response

time. For the development of a graphical user interface in network-based

information system for use in multi-institutional clinical trials, HTTP/HTML has

several advantages over an ordinary client-server model. Therefore, we concluded

that we would adopt HTP/HTML for the construction of user interfaces for

physicians in each spital and for data managers in our coordinating center.

PMID: 8974208 [Indexed for MEDLINE]

3902. J Fla Med Assoc. 1996 Nov;83(9):634-8.

The Internet & Healthcare Education: HELIX.

Roy RT(1), Merril JR.

Author information:

(1)Medical Consumer Media, Bethesda, Maryland, USA.

With the advent of the World Wide Web (WWW), we are now on the cusp of a

revolution in computer technology that will dramatically enhance medical

education. An historical analogy might be Johann Gutenberg's invention of movable

type in the 1400's-radically decreasing the cost, time, and expertise required to

reproduce printed materials. Now, the WWW can decrease the cost of disseminating

medical educational materials. When an educational module is authored for the

Web, it can be placed on a computer "server" which in turn, distributes the

program on the WWW to anyone with a computer and Internet access. Rapidly

emerging standards are being developed to allow increasingly rich educational

experiences on the Internet. With the introduction of HTML (hypertext markup

language), a standardized method of placing text and graphics, as well as the

connections between them, was created.

PMID: 9160009 [Indexed for MEDLINE]

3903. Protein Sci. 1996 Nov;5(11):2203-16.

Cleavage site analysis in picornaviral polyproteins: discovering cellular targets

by neural networks.

Blom N(1), Hansen J, Blaas D, Brunak S.

Author information:

(1)Center for Biological Sequence Analysis, Technical University of Denmark.

Picornaviral proteinases are responsible for maturation cleavages of the viral

polyprotein, but also catalyze the degradation of cellular targets. Using

graphical visualization techniques and neural network algorithms, we have

investigated the sequence specificity of the two proteinases 2Apro and 3Cpro. The

cleavage of VP0 (giving rise to VP2 and VP4), which is carried out by a so-far

unknown proteinase, was also examined. In combination with a novel surface

exposure prediction algorithm, our neural network approach successfully

distinguishes known cleavage sites from noncleavage sites and yields a more

consistent definition of features common to these sites. The method is able to

predict experimentally determined cleavage sites in cellular proteins. We present

a list of mammalian and other proteins that are predicted to be possible targets

for the viral proteinases. Whether these proteins are indeed cleaved awaits

experimental verification. Additionally, we report several errors detected in the

protein databases. A computer server for prediction of cleavage sites by

picornaviral proteinases is publicly available at the e-mail address

NetPicoRNA@cbs.dtu.dk or via WWW at http:@www.cbs.dtu.dk/services/NetPicoRNA/.

DOI: 10.1002/pro.5560051107

PMCID: PMC2143287

PMID: 8931139 [Indexed for MEDLINE]

3904. J Am Med Inform Assoc. 1996 Sep-Oct;3(5):318-27.

Remote analysis of physiological data from neurosurgical ICU patients.

Nenov V(1), Klopp J.

Author information:

(1)Division of Neurosurgery, and Brain Research Institute, UCLA School of

Medicine 90095, USA. nenov@neurosurg.medsch.ucla.edu

Erratum in

J Am Med Inform Assoc 1997 Jan-Feb;4(1):70.

Recent technical advances in Internet-based client/server applications and new

multimedia communications protocols are enabling the development of

cost-effective, platform-independent solutions to the problem of remote access to

continuously acquired physiological data. The UCLA Neurosurgery Intensive Care

Unit (ICU) has developed a distributed computer system that provides access over

the World Wide Web (WWW) to current and previously acquired physiological data,

such as intracranial pressure, cerebral perfusion pressure, and heart rate from

critical care patients. Physicians and clinical researchers can access these data

through personal computers from their offices, from their homes, or even while on

the road. The system creates and continuously updates a database of all monitored

parameters in data formats that can readily be used for further clinical studies.

This paper describes an extension to this system that allows for remote

interaction with and analysis of the data via the WWW. Physicians can now pose a

limited, predefined set of clinically relevant questions to the system without

having to be at the patient's bedside.

PMCID: PMC116316

PMID: 8880679 [Indexed for MEDLINE]

3905. Yale J Biol Med. 1996 Sep-Oct;69(5):439-44.

Development of an academic Internet resource.

Ruskin KJ(1), Doyle DJ, Engel TP.

Author information:

(1)Department of Anesthesiology, Yale University School of Medicine, New Haven,

Connecticut 06520, USA. ruskin@gasnet.med.yale.edu

Networked electronic publication is a relatively new development that has already

begun to change the way in which medical information is exchanged. Electronic

publications can present ideas that would be impossible in printed text, using

multimedia components such as sound and movies. Physicians who use the World Wide

Web (WWW) on a regular basis may recognize the value of electronic publication

and decide to become information providers. Nearly anyone with a computer and

modem can create a WWW resource on a Web server at a hospital or on a commercial

hosting service. Medical publication on the Internet demands a high level of

quality control because the information will be available to anyone who cares to

look. Creating a peer-review system for electronic information may, therefore,

help to enhance academic recognition of Internet medical resources. Resources

containing medical information must be continually available and protected from

system failures and unauthorized access. As Internet technology matures and these

problems are solved, electronic publication may become the predominant method of

communication between medical professionals.

PMCID: PMC2588996

PMID: 9381739 [Indexed for MEDLINE]

3906. Comput Appl Biosci. 1996 Aug;12(4):297-301.

GEOMETRY: a software package for nucleotide sequence analysis using statistical

geometry in sequence space.

Kuznetsov I(1), Morozov P.

Author information:

(1)Laboratory of Molecular Evolution, Institute of Cytology and Genetics,

Novosibirsk, Russia. kuznets@benpc.bionet.nsk.su

GEOMETRY is a software package for the analysis of nucleotide sequences using the

method of statistical geometry in sequence space. The package consists of

programs performing estimation of the average geometry of sequence quartets,

analysis of positional variability and computer simulation of parallel and

tree-like sequence divergence with user-defined parameters. It provides an

independent tool for evaluation of the reliability of conventional phylogenetic

trees and calibration of the time of sequence divergence. GEOMETRY may be of

interest for all scientists engaged in the study of molecular phylogeny. The

package is available by anonymous FTP from ftp.bionet.nsk.su, directory

/incoming/molevol/geom.exe, and will be available from EMBL file server (URL:

http://@www.ebi.ac.uk)

PMID: 8902356 [Indexed for MEDLINE]

3907. Radiol Med. 1996 Jul-Aug;92(1-2):122-7.

[A computer system for the systematization of MR findings in knee joint

diseases].

[Article in Italian]

Sparacia G(1), Lo Casto A, Mercurio G, Brancato M, Bartolotta T, Lagalla R.

Author information:

(1)Istituto di Radiologia P. Cignolini, Università di Palermo.

sparacia@cere.pa.cnr.it

An inexpensive, easy-access computer-based system is proposed, which was

developed for the systematization of the clinical series of knee joint disorders

studied with Magnetic Resonance Imaging (MRI). The system is based upon the

integration of multimedia technology and Data Base Management Systems (DBMS). The

hardware configuration for this project included an Apple Macintosh workstation

based on a Motorola 68040 microprocessor and a customized application developed

by the authors with the 4th Dimension software. The MR images available only on

film were digitized off-line with a solid-state Charge Coupled Device (CCD)

scanner with back-light cover for transparency. Otherwise, MR images were

acquired on-line through an Ethernet-based local area network from the MR unit or

from a SparcStation-Advantage Windows workstation connected with the MR unit.

Image post-processing was performed with the Adobe PhotoShop software. The system

was devoted to the systematization and analysis of a clinical series of 800 MR

studies of the knee. A mean of 10 significant MR images were stored for each

examination with a standard image compression algorithm--the Joint Photographic

Experts Group (JPEG). This permitted us to save the system's storage space and at

the same time to preserve image quality for consultation and teaching purposes,

not for diagnosis which is made on the backboard or on the MR unit's or Advantage

Window's monitor. Finally, MR findings were indexed with a customized check-list

specific for knee joint disorders. On the basis of stored and selected

information, it was thus possible to carry out a statistical analysis and to make

detailed reports which are useful for scientific purposes, such as the

preparation of lectures and papers. Moreover, the system was very useful for

patients' follow-up and for the preparation of hypermedia teaching applications

on knee joint disorders which are available on the Internet at our World-Wide Web

server (URL: http://mbox.unipa.it/radpa/radpa ++ +.html). In conclusion, the

system is a cost-effective and user-friendly solution for the multipurpose

management of radiologic series.

PMID: 8966250 [Indexed for MEDLINE]

3908. J Mol Biol. 1996 Jun 21;259(4):840-54.

FASTA-SWAP and FASTA-PAT: pattern database searches using combinations of aligned

amino acids, and a novel scoring theory.

Ladunga I(1), Wiese BA, Smith RF.

Author information:

(1)Department of Molecular and Human Genetics, Baylor College of Medicine,

Houston, TX 77030, USA.

We introduce two new pattern database search tools that utilize statistical

significance and information theory to improve protein function identification.

Both the general pattern scoring theory with the specific matrices introduced

here and the low redundancy of pattern databases increase search sensitivity and

selectivity. Pattern scoring preferentially rewards matches at conserved

positions in a pattern with higher scores than matches at variable positions, and

assigns more negative scores to mismatches at conserved positions than to

mismatches at variable positions. The theory of pattern scoring can be used to

create log-odds pattern scores for patterns derived from any set of multiple

alignments. This theoretical framework can be used to adapt existing sequence

database search tools to pattern analysis. Our FASTA-SWAP and FASTA-PAT tools are

extensions of the FASTA program that search a sequence query against a pattern

database. In the first step, FASTA-SWAP searches the diagonals of the query

sequence and the library pattern for high-scoring segments, while FASTA-PAT

performs an extended version of hashing. In the second step, both methods refine

the alignments and the scores using dynamic programming. The tools utilize an

extremely compact binary representation of all possible combinations of amino

acid residues in aligned positions. Our FASTA-SWAP and FASTA-PAT tools are well

suited for functional identification of distant relatives that may be missed by

sequence database search methods. FASTA-SWAP and FASTA-PAT searches can be

performed using our World-Wide Web Server

(http://dot.imgen.bcm.tmc.edu:9331/seq-search/Op tions/fastapat.html).

DOI: 10.1006/jmbi.1996.0362

PMID: 8683587 [Indexed for MEDLINE]

3909. Gene. 1996 Jun 12;172(1):GC43-50.

Adapting EcoCyc for use on the World Wide Web.

Paley SM(1), Karp PD.

Author information:

(1)Artificial Intelligence Center, SRI International, Menlo Park, CA 94025, USA.

paley@ai.sri.com

The World Wide Web (WWW) offers the potential to deliver specialized information

to an audience of unprecedented size. Along with this exciting new opportunity

comes a challenge for software developers: instead of rewriting our software

applications to operate over the WWW, how can we maximize software reuse by

retrofitting existing applications? We have developed a Web server tool, written

in Common Lisp, that allows existing graphical user interface applications

written using the Common Lisp Interface Manager (CLIM) to hook easily into the

WWW. This tool-CWEST (CLIM-WEb Server Tool, pronounced "quest")-was developed to

operate with EcoCyc, an electronic encyclopedia of the genes and metabolism of

the bacterium E. coli. EcoCyc consists of a database of objects relevant to E.

coli biochemistry and a user interface, implemented in CLIM, that runs on the

X-window system and generates graphical displays appropriate to biological

objects. Each query to the EcoCyc WWW server is treated as a command to the

EcoCyc program, which dynamically generates an appropriate CLIM drawing. CWEST

translates that drawing, which can be a mixture of text and graphics, into the

HyperText Markup Language (HTML) and/or the Graphics Interchange Format (GIF),

which are returned to the client. Sensitive regions embedded in the CLIM drawing

are converted to hyperlinks with Universal Resource Locators (URLs) that generate

further EcoCyc queries. This tight coupling of CLIM output with Web output makes

CLIM a powerful high-level programming tool for Web applications. The flexibility

of Common Lisp and CLIM made implementation of the server tool surprisingly easy,

requiring few changes to the existing EcoCyc program. The results can be seen at

URL http: @www.ai.sri.com/ecocyc/browser.html. We have made CWEST available to

the CLIM community at large, with the hope that it will spur other software

developers to make their CLIM applications available over the WWW.

PMID: 8654966 [Indexed for MEDLINE]

3910. AJR Am J Roentgenol. 1996 Jun;166(6):1265-7.

Efficient searching for specific resources on the World-Wide Web: creation of a

search server for radiologists.

Weinberger E(1), Manber U.

Author information:

(1)Department of Radiology, Children's Hospital and Medical Center, Seattle, WA

98105, USA.

OBJECTIVE: Our objective was to develop an efficient way to search for radiology

information on the World-Wide Web (WWW).

MATERIALS AND METHODS: With the Harvest Information Discovery and Access System

and a DECstation 3000 computer, each week we gather and update WWW documents that

relate to radiology. To limit the size of the gathered documents, certain data

(such as image files) are omitted. To date, we have gathered from 20 main

radiology sites. By accessing our server, thousands of documents can be

efficiently searched for specific radiology information.

RESULTS: At no charge, individuals with access to the WWW can request a search of

our server for specific radiology information. We return to the requester in

HyperText Markup Language a document with results of the search included as

embedded pointers. Our server address, or uniform resource locator, is

http://glimpse.cs.arizona.edu/radiology.html.

CONCLUSION: Use of a server dedicated to radiology is an efficient way to search

the internet for radiology resources, such as teaching files. We provide access

to many sites with one search, and we select those sites to ensure that only

relevant and timely information is gathered. We also index that information.

DOI: 10.2214/ajr.166.6.8633428

PMID: 8633428 [Indexed for MEDLINE]

3911. Comput Appl Biosci. 1996 Jun;12(3):237-40.

The directory of P450-containing systems on WorldWide Web.

Degtyarenko KN(1), Fábián P.

Author information:

(1)International Centre for Genetic Engineering and Biotechnology, Trieste,

Italy. kirill@icgeb.trieste.it

To facilitate access to electronic resources for all researchers working in the

field of P450 proteins and P450-containing systems, a WorldWide Web server has

been established called The Directory of P450-containing Systems at <

http://www.icgeb.trieste.it/p450/ >. Currently it contains the most up-to-date

list of sequences of both the P450 superfamily and proteins mediating electron

transfer to P450, i.e. NADPH:P450 reductases, specific NAD(P)H:ferredoxin

reductases, cytochrome b5 reductases, ferredoxins and cytochromes b5, and their

homologues from different enzyme systems. All the referenced sequences are

provided with accession numbers and cross-links to major sequence databanks: PIR,

SWISS-PROT, EMBL/GenBank and PRF.

PMID: 8872393 [Indexed for MEDLINE]

3912. Comput Appl Biosci. 1996 Jun;12(3):231-5.

Development of an animal genome database and its search system.

Wada Y(1), Yasue H.

Author information:

(1)Animal Genome Research Group, National Institute of Animal Industry, Ibaraki,

Japan.

An animal genome database has been developed on a Unix workstation and maintained

by a relational database management system. This database has focused on the

comparative gene mapping between species to assist the mapping of the genes

related to phenotypic traits in livestock. The linkage maps, cytogenetic maps,

polymerase chain reaction primers of pig, cattle, mouse and human, and their

references have been included in the database, and the correspondence among

species have been stipulated in the database. In order to search the database

effectively, the World Wide Web server (http://ws4.niai.affrc.go.jp/) and the

electronic mail server system (e-mail: jgbase-mail@ niai.affrc.go.jp) have been

developed on different Unix workstations. These servers are connected to the

Internet.

PMID: 8872392 [Indexed for MEDLINE]

3913. Genome Res. 1996 May;6(5):454-62.

BCM Search Launcher--an integrated interface to molecular biology data base

search and analysis services available on the World Wide Web.

Smith RF(1), Wiese BA, Wojzynski MK, Davison DB, Worley KC.

Author information:

(1)Department of Molecular and Human Genetics, Baylor College of Medicine,

Houston, Texas 77030, USA. rsmith@bcm.tmc.edu

The BCM Search Launcher is an integrated set of World Wide Web (WWW) pages that

organize molecular biology-related search and analysis services available on the

WWW by function, and provide a single point of entry for related searches. The

Protein Sequence Search Page, for example, provides a single sequence entry form

for submitting sequences to WWW servers that offer remote access to a variety of

different protein sequence search tools, including BLAST, FASTA, Smith-Waterman,

BEAUTY, PROSITE, and BLOCKS searches. Other Launch pages provide access to (1)

nucleic acid sequence searches, (2) multiple and pair-wise sequence alignments,

(3) gene feature searches, (4) protein secondary structure prediction, and (5)

miscellaneous sequence utilities (e.g., six-frame translation). The BCM Search

Launcher also provides a mechanism to extend the utility of other WWW services by

adding supplementary hypertext links to results returned by remote servers. For

example, links to the NCBI's Entrez data base and to the Sequence Retrieval

System (SRS) are added to search results returned by the NCBI's WWW BLAST server.

These links provide easy access to auxiliary information, such as Medline

abstracts, that can be extremely helpful when analyzing BLAST data base hits. For

new or infrequent users of sequence data base search tools, we have preset the

default search parameters to provide the most informative first-pass sequence

analysis possible. We have also developed a batch client interface for Unix and

Macintosh computers that allows multiple input sequences to be searched

automatically as a background task, with the results returned as individual HTML

documents directly to the user's system. The BCM Search Launcher and batch client

are available on the WWW at URL http:@gc.bcm.tmc.edu:8088/search-launcher.html.

PMID: 8743995 [Indexed for MEDLINE]

3914. Radiographics. 1996 May;16(3):683-91.

Structured entry of radiology reports using World Wide Web technology.

Kahn CE Jr(1), Wang K, Bell DS.

Author information:

(1)Department of Radiology, Medical College of Wisconsin, Milwaukee 53226, USA.

Structured data entry--in which information is entered by using predetermined

data elements and formats--has the potential to improve the radiology reporting

process. The dependence on particular computer hardware and software platforms

has posed a barrier to wider use of this approach. The World Wide Web (WWW), a

client-server protocol for delivery of multimedia data via the Internet, was used

to achieve platform-independent structured entry of radiology reports. A

developmental system for structured entry of radiology reports, called SPIDER,

incorporates a knowledge base of hierarchically organized concepts, a WWW server,

and two specialized programs. The WebForm program transforms the system's

knowledge into graphical WWW data-entry forms; the WebReport program converts

data entered on these forms into outline-format reports. SPIDER received

favorable evaluations from sonographers and physicians who used the system to

record the results of several test cases. WWW technology can be used to achieve

platform-independent entry of the results of radiologic procedures.

DOI: 10.1148/radiographics.16.3.8897632

PMID: 8897632 [Indexed for MEDLINE]

3915. Radiol Med. 1996 May;91(5):622-6.

[Implementation of a server World Wide Web of radiology accessible by Internet].

[Article in Italian]

Caramella D(1), Neri E, Del Sarto M, Lencioni R, Bartolozzi C.

Author information:

(1)Cattedra di Radiologia, Università di Pisa.

Internet is an international computer network that uses standard communication

protocols for the exchange of information. This facilitates the retrieval of

multimedia data through a "web" of servers distributed in the whole world. Among

Internet users, Radiologists are a potentially important segment, due to the

inherent multimedia characteristics of the discipline, which requires a

continuous international update of information. The Department of Radiology of

the University of Pisa has an Internet access through the metropolitan area

network which was installed in the framework of the CNR Telecomunicazioni

Project. The Internet access allowed the implementation of a World Wide Web

server made public on Internet in March, 1994, being the first European server

specifically oriented to radiology. This server can be accessed at the following

address: http:@www.rad.unipi.it:7080/IRMosaicHome.html . On the server, 3

hypermedia papers are present, a list of international servers containing

radiological information, a questionnaire, and statistics concerning the number

of users who accessed the server. In the first 18 months of public access through

Internet (April 1, 1994-September 30, 1995) 16,166 users accessed the server,

retrieving 127,349 documents, corresponding to 1,279.7 MByte of information. The

mean amount of information retrieved in each access to the server in the

considered quarters ranges from 72 to 85.8 kByte. The geographic distribution of

the users who accessed the server is the following: United States, 7,158, Italy,

2,466; other European countries, 3,813; other extra-European countries, 2,729.

The increasingly diffuse knowledge of Internet services had a substantial impact

on the rise in number of the servers and of the users who can access them. It is

likely that in the future this technology will be used with increasing interest

by Radiologists, since it provides easier "navigation" through multimedia

information, consisting of text and several images, without the inherent

limitations of the printed paper.

PMID: 8693130 [Indexed for MEDLINE]

3916. Comput Appl Biosci. 1996 Apr;12(2):151-5.

A Tcl-based SRS v. 4 interface.

Schaftenaar G(1), Cuelenaere K, Noordik JH, Etzold T.

Author information:

(1)CAOS/CAMM Center, Faculty of Science, Nijmegen University, The Netherlands.

A new SRS (Sequence Retrieval System) user interface has been developed for SRS

v.4. Key features are the support of simple character-oriented (ASCII, VT100)

terminals by coding in Tcl augmented by some dedicated Curses calls, support of

graphics terminals in an X-Windows version by using the Tk extension to Tcl, and

support of a client/server environment by using the TDP extension to Tcl. The

Sequence Retrieval System (SRS) is a powerful tool for the fast extraction of

information from flat file libraries (Etzold and Argos, 1993) and has rapidly

established itself as a major research instrument for the bio-informatics

community. Internally the system employs a query language, which is user

accessible through either a command-line user interface, 'getz', or a more user

friendly, character-oriented window interface. For SRS versions up to release v.

3, this window interface supported VT100-compatible terminals. Because of major

changes in the underlying SRS libraries, the v. 3 interface became fully

incompatible with the most recent version of SRS (v. 4.x). Thus the many users

with only a simple terminal/terminal emulator connection were either deprived of

access to SRS, or were forced to use the ASCII WWW client LYNX. This prompted us

to develop a character-oriented SRS v. 4 window interface with the look and feel

of its SRS v. 3.1 predecessor and coded to be as library independent as possible

to maintain compatibility with future SRS releases. In addition, some

'extensions' were coded to widen the applicability to graphics terminals and to a

client/server environment. At the time of preparation of this paper, the SRS

interface described had been implemented in one form or another on most EM Bnet

nodes and on all the platforms given in Table II. The code has been stored at the

EMBL in Heidelberg, where it will be available, with installation instructions

and scripts, as part of the SRS distribution.

PMID: 8744778 [Indexed for MEDLINE]

3917. Electrophoresis. 1996 Mar;17(3):566-72.

A World Wide Web (WWW) server database engine for an organelle database, MitoDat.

Lemkin PF(1), Chipperfield M, Merril C, Zullo S.

Author information:

(1)Image Processing Section/LMMB, NCI-FCRDC/NIH, Frederick, MD 21702, USA.

lemkin/ncifcrf.gov

We describe a simple database search engine "dbEngine" which may be used to

quickly create a searchable database on a World Wide Web (WWW) server. Data may

be prepared from spreadsheet programs (such as Excel, etc.) or from tables

exported from relationship database systems. This Common Gateway Interface

(CGI-BIN) program is used with a WWW server such as available commercially, or

from National Center for Supercomputer Algorithms (NCSA) or CERN. Its

capabilities include: (i) searching records by combinations of terms connected

with ANDs or ORs; (ii) returning search results as hypertext links to other WWW

database servers; (iii) mapping lists of literature reference identifiers to the

full references; (iv) creating bidirectional hypertext links between pictures and

the database. DbEngine has been used to support the MitoDat database (Mendelian

and non-Mendelian inheritance associated with the Mitochondrion) on the WWW.

DOI: 10.1002/elps.1150170327

PMID: 8740181 [Indexed for MEDLINE]

3918. Electrophoresis. 1996 Mar;17(3):556-65.

The yeast SWISS-2DPAGE database.

Sanchez JC(1), Golaz O, Frutiger S, Schaller D, Appel RD, Bairoch A, Hughes GJ,

Hochstrasser DF.

Author information:

(1)Clinical Chemistry Laboratory, Geneva Hospital, Switzerland.

sanchez/dim.hcuge.CH

The systematic sequencing of the yeast genome will soon be completed. A new

challenge has been launched by the EUROFAN (European Functional Analysis) project

whose goal is to elucidate the physiological and biochemical function of newly

discovered open reading frames (ORF) from yeast. One of the approaches is to use

protein-based technologies such as two-dimensional gel electrophoresis and

protein identification in order to establish a yeast reference map. Modified

protein patterns can be compared to the reference map which hopefully will help

identify changes related, for example, to growth processes or developmental

events. This paper describes the yeast SWISS-2DPAGE database in which charge

separation was obtained using immobilized pH gradient (IPG). Proteins identified

by gel comparison, amino acid composition analysis and/or microsequencing are

recorded and described in an accessible uniform format. We have identified more

than one hundred polypeptides, several of which were newly mapped. In addition,

the yeast SWISS-2DPAGE database can be freely accessed through the World Wide Web

(WWW) network on the ExPASy molecular biology server.

DOI: 10.1002/elps.1150170326

PMID: 8740180 [Indexed for MEDLINE]

3919. Gerontologist. 1996 Feb;36(1):100-5.

Gero-informatics and the Internet: locating gerontology information on the World

Wide Web (WWW).

Ellis RD(1), Jankowski TB, Jasper JE, Abdul A.

Author information:

(1)Institute of Gerontology, Wayne State University, Detroit, MI 48202, USA.

The Internet, and in particular the World Wide Web (WWW), are growing at a

tremendous rate. Information useful to researchers and practitioners in

gerontology is scattered and hard to locate, even for experts in the use of the

WWW. In order to locate and organize access to this material, a WWW server was

developed for gerontologists, using state-of-the-art Internet search techniques.

This report provides background on the WWW, reasons for its growth, its potential

usefulness to gerontologists, and the results of an exhaustive search of over 300

potential sites. Relevant information was discovered in 5 general categories of

gerontology-related information: academic institutions, government agencies,

biomedical and health research institutions, general interest sites and data

archives.

PMID: 8932415 [Indexed for MEDLINE]

3920. Biochimie. 1996;78(5):364-9.

WWW-query: an on-line retrieval system for biological sequence banks.

Perrière G(1), Gouy M.

Author information:

(1)Laboratoire de Biométrie, Génétique et Biologie des Populations, Université

Claude-Bernard Lyon-1, Villeurbanne, France.

We have developed a World Wide Web (WWW) version of the sequence retrieval system

Query: WWW-Query. This server allows to query nucleotide sequence banks in the

EMBL/GenBank/DDBJ formats and protein sequence banks in the NBRF/PIR format.

WWW-Query includes all the features of the on-line sequences browsers already

available: possibility to build complex queries, integration of cross-references

with different data banks, and access to the functional zones of biological

interest. It also provides original services not available elsewhere:

introduction of the notion of re-usable sequence lists, integration of dedicated

helper applications for visualizing alignments and phylogenetic trees and links

with multivariate methods for studying codon usage or for complementing

phylogenies.

PMID: 8905155 [Indexed for MEDLINE]

3921. Eisei Shikenjo Hokoku. 1996;(114):76-83.

[An international exchange and dissemination of chemical safety information on

the Internet].

[Article in Japanese]

Ohtake C, Yamamoto M, Nakano T, Nakata K, Ishikawa K, Kaminuma T.

An information system for chemical safety has been developed on the National

Institute of Health Sciences (NIHS) Information and Computing Infrastructure. The

system is based on client server systems on the local area network (LAN)

connected to the Internet. A wide range of safety information for chemicals

including foods, food additives, household goods, industrial chemicals and

environmental pollutants were collected and put on the World Wide Web (WWW)

server and the database management system, Sybase. In addition to original

information contents, the System has links to many useful Web sites so that it

functions as a global hub for chemical safety information.

PMID: 9037870 [Indexed for MEDLINE]

3922. Nucleic Acids Res. 1996 Jan 1;24(1):82-5.

The Ribosomal Database Project (RDP).

Maidak BL(1), Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, Woese CR.

Author information:

(1)Department of Microbiology, University of Illinois, Urbana 61801-3704, USA.

The Ribosomal Database Project (RDP) is a curated database that offers

ribosome-related data, analysis services and associated computer programs. The

offerings include phylogenetically ordered alignments of ribosomal RNA (rRNA)

sequences, derived phylogenetic trees, rRNA secondary structure diagrams and

various software for handling, analyzing and displaying alignments and trees. The

data are available via anonymous ftp (rdp.life.uiuc.edu), electronic mail

(server@rdp.life.uiuc.edu), gopher (rdpgopher.life.uiuc.edu) and World Wide Web

(WWW)(http://rdpwww.life.uiuc.edu/). The electronic mail and WWW servers provide

ribosomal probe checking, screening for possible chimeric rRNA sequences,

automated alignment and approximate phylogenetic placement of user-submitted

sequences on an existing phylogenetic tree.

PMCID: PMC145599

PMID: 8594608 [Indexed for MEDLINE]

3923. Nucleic Acids Res. 1996 Jan 1;24(1):248-52.

O-GLYCBASE: a revised database of O-glycosylated proteins.

Hansen JE(1), Lund O, Nielsen JO, Hansen JE, Brunak S.

Author information:

(1)Laboratory for Infectious Diseases, University of Copenhagen, Denmark.

O-GLYCBASE is a comprehensive database of information on glycoproteins and their

O-linked glycosylation sites. Entries are compiled and revised from the

SWISS-PROT and PIR databases as well as directly from recently published reports.

Nineteen percent of the entries extracted from the databases needed revision with

respect to O-linked glycosylation. Entries include information about species,

sequence, glycosylation site and glycan type, and are fully referenced. Sequence

logos displaying the acceptor specificity for the GaINAc transferase are shown. A

neural network method for prediction of mucin type O-glycosylation sites in

mammalian glycoproteins exclusively from the primary sequence is made available

by E-mail or WWW. The O-GLYCBASE database is also available electronically

through our WWW server or by anonymous FTP.

PMCID: PMC145605

PMID: 8594592 [Indexed for MEDLINE]

3924. Nucleic Acids Res. 1996 Jan 1;24(1):221-2.

The ENZYME data bank in 1995.

Bairoch A(1).

Author information:

(1)Department of Medical Biochemistry, University of Geneva, Switzerland.

The ENZYME data bank is a repository of information relative to the nomenclature

of enzymes. The current version (October 1995) contains information relevant to

3594 enzymes. It is available from a variety of file and ftp servers as well as

through the ExPASy World Wide Web server (http://expasy.hcuge.ch/).

PMCID: PMC145615

PMID: 8594586 [Indexed for MEDLINE]

3925. Nucleic Acids Res. 1996 Jan 1;24(1):210-3.

The SBASE protein domain library, Release 4.0: a collection of annotated protein

sequence segments.

Murvai J(1), Gabrielian A, Fábián P, Hátsagi Z, Degtyarenko K, Hegyi H, Pongor S.

Author information:

(1)ABC Institute for Biochemistry and Protein Research, Gödöllö, Hungary.

SBASE 4.0 is the fourth release of SBASE, a collection of annotated protein

domain sequences that represent various structural, functional, ligand binding

and topogenic segments of proteins. SBASE was designed to facilitate the

detection of functional homologies and can be searched with standard database

search tools, such as FASTA and BLAST3. The present release contains 61 137

entries provided with standardized names and cross-referenced to all major

protein, nucleic acid and sequence pattern collections. The entries are clustered

into 13 155 groups in order to facilitate detection of distant similarities.

SBASE 4.0 is freely available by anonymous ftp file transfer from

ftp.icgeb.trieste.it. Individual records can be retrieved with the gopher server

at icgeb.trieste.it and with a World Wide Web server at

http://www.icgeb.trieste.it. Automated searching of SBASE with BLAST can be

carried out with the electronic mail server sbase@icgeb.trieste.it, which now

also provides a graphic representation of the homologies. A related mail server,

domain@hubi.abc.hu, assigns SBASE domain homologies on the basis of SWISS-PROT

searches.

PMCID: PMC145610

PMID: 8594582 [Indexed for MEDLINE]

3926. Nucleic Acids Res. 1996 Jan 1;24(1):182-8.

Progress with the PRINTS protein fingerprint database.

Attwood TK(1), Beck ME, Bleasby AJ, Degtyarenko K, Parry Smith DJ.

Author information:

(1)Department of Biochemistry, University College London, UK.

PRINTS is a compendium of protein motif 'fingerprints' derived from the OWL

composite sequence database. Fingerprints are groups of motifs within sequence

alignments whose conserved nature allows them to be used as signatures of family

membership. To date, 400 fingerprints have been constructed and stored in Prints,

the size of which has doubled in the last year. The current version, 9.0, encodes

approximately 2000 motifs, covering a range of globular and membrane proteins,

modular polypeptides, and so on. Fingerprints inherently offer improved

diagnostic reliability over single motif methods by virtue of the mutual context

provided by motif neighbours. PRINTS thus provides a useful adjunct to the widely

used PROSITE dictionary of patterns. The database is now accessible via the

Database Browser on the UCL Bioinformatics server at

http://www.biochem.ucl.ac.uk/bsm/dbbrowser .

PMCID: PMC145564

PMID: 8594576 [Indexed for MEDLINE]

3927. Nucleic Acids Res. 1996 Jan 1;24(1):132-6.

A cholinesterase genes server (ESTHER): a database of cholinesterase-related

sequences for multiple alignments, phylogenetic relationships, mutations and

structural data retrieval.

Cousin X(1), Hotelier T, Liévin P, Toutant JP, Chatonnet A.

Author information:

(1)Unité des Venins, Institut Pasteur, Paris, France.

We have built a database of sequences phylogenetically related to cholinesterases

(ESTHER) for esterases, alpha/beta hydrolase enzymes and relatives). These

sequences define a homogeneous group of enzymes (carboxylesterases, lipases and

hormone-sensitive lipases) with some related proteins devoid of enzymatic

activity. The purpose of ESTHER is to help comparison and alignment of any new

sequence appearing in the field, to favour mutation analysis of

structure-function relationships and to allow structural data recovery. ESTHER is

a World Wide Web server with the URL

http://www.montpellier.inra.fr:70/cholinesterase.

PMCID: PMC145568

PMID: 8594562 [Indexed for MEDLINE]

3928. Proc AMIA Annu Fall Symp. 1996:757-61.

Lamprey: tracking users on the World Wide Web.

Felciano RM(1), Altman RB.

Author information:

(1)Section on Medical Informatics, Stanford University, California, USA.

felciano@smi.stanford.edu

Tracking individual web sessions provides valuable information about user

behavior. This information can be used for general purpose evaluation of

web-based user interfaces to biomedical information systems. To this end, we have

developed Lamprey, a tool for doing quantitative and qualitative analysis of

Web-based user interfaces. Lamprey can be used from any conforming browser, and

does not require modification of server or client software. By rerouting WWW

navigation through a centralized filter, Lamprey collects the sequence and timing

of hyperlinks used by individual users to move through the web. Instead of

providing marginal statistics, it retains the full information required to

recreate a user session. We have built Lamprey as a standard Common Gateway

Interface (CGI) that works with all standard WWW browsers and servers. In this

paper, we describe Lamprey and provide a short demonstration of this approach for

evaluating web usage patterns.

PMCID: PMC2233185

PMID: 8947767 [Indexed for MEDLINE]

3929. Proc AMIA Annu Fall Symp. 1996:733-7.

Implications of the Java language on computer-based patient records.

Pollard D(1), Kucharz E, Hammond WE.

Author information:

(1)Division of Medical Informatics, Duke University Medical Center, Durham, NC,

USA.

The growth of the utilization of the World Wide Web (WWW) as a medium for the

delivery of computer-based patient records (CBPR) has created a new paradigm in

which clinical information may be delivered. Until recently the authoring tools

and environment for application development on the WWW have been limited to Hyper

Text Markup Language (HTML) utilizing common gateway interface scripts. While, at

times, this provides an effective medium for the delivery of CBPR, it is a less

than optimal solution. The server-centric dynamics and low levels of

interactivity do not provide for a robust application which is required in a

clinical environment. The emergence of Sun Microsystems' Java language is a

solution to the problem. In this paper we examine the Java language and its

implications to the CBPR. A quantitative and qualitative assessment was

performed. The Java environment is compared to HTML and Telnet CBPR environments.

Qualitative comparisons include level of interactivity, server load, client load,

ease of use, and application capabilities. Quantitative comparisons include data

transfer time delays. The Java language has demonstrated promise for delivering

CBPRs.

PMCID: PMC2232959

PMID: 8947762 [Indexed for MEDLINE]

3930. Proc AMIA Annu Fall Symp. 1996:669-73.

The "SentiWeb" method for exploring a database on the Net.

Boussard E(1), Flahault A.

Author information:

(1)INSERM U444, Institut fédératif Saint-Antoine de Recherche sur la Santé,

Paris, France.

Return of information is one of the main goals of any public health information

system. About 25,000 maps and 10,000 graphs may be obtained from the time-series

collected in the database of the French Communicable Diseases Network (FCDN).

Furthermore, this huge epidemiological atlas is updated each week. What is the

optimal way of returning such information? This report discloses the strategies

used for enhancing the access facilities to the FCDN database for any users,

particularly those without specific training in epidemiology or database query

language. The technical options implemented in the SentiWeb server

(http:/(/)www.b3e.jussieu.fr) are discussed.

PMCID: PMC2233219

PMID: 8947749 [Indexed for MEDLINE]

3931. Proc AMIA Annu Fall Symp. 1996:348-52.

Integrating DXplain into a clinical information system using the World Wide Web.

Elhanan G(1), Socratous SA, Cimino JJ.

Author information:

(1)Department of Medical Informatics, Columbia Presbyterian Medical Center, New

York, New York, USA.

The World Wide Web(WWW) offers a cross-platform environment and standard

protocols that enable integration of various applications available on the

Internet. The authors use the Web to facilitate interaction between their

Web-based Clinical Information System and a decision-support system-DXplain, at

the Massachusetts General Hospital-using local architecture and Common Gateway

Interface programs. The current application translates patients laboratory test

results into DXplain's terms to generate diagnostic hypotheses. Two different

access methods are utilized for this model; Hypertext Transfer Protocol (HTTP)

and TCP/IP function calls. While clinical aspects cannot be evaluated as yet, the

model demonstrates the potential of Web-based applications for interaction and

integration and how local architecture, with a controlled vocabulary server, can

further facilitate such integration. This model serves to demonstrate some of the

limitations of the current WWW technology and identifies issues such as control

over Web resources and their utilization and liability issues as possible

obstacles for further integration.

PMCID: PMC2233176

PMID: 8947686 [Indexed for MEDLINE]

3932. Proc AMIA Annu Fall Symp. 1996:179-83.

A Web-based architecture for the intelligent management of chronic patients.

Riva A(1).

Author information:

(1)Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy.

We describe a distributed architecture for medical informatics applications,

based on the World-Wide Web (WWW) environment. After discussing previous

experiences in the application of the WWW for medical purposes, we outline the

features of a Common Lisp HTTP server designed to provide access to medical

informatics applications using a standard Web browser. As an example of

application, we describe a system for therapy planning and revision in the field

of insulin-dependent diabetes. The system performs automatic data analysis and

interpretation and provides advice on possible adjustments to the therapeutic

protocol that the patients are following, taking advantage of the network and

multimedia capabilities offered by the WWW for user interaction.

PMCID: PMC2233199

PMID: 8947652 [Indexed for MEDLINE]

3933. Receptors Channels. 1996;4(3):161-4.

Automated modelling of the transmembrane region of G-protein coupled receptor by

Swiss-model.

Peitsch MC(1), Herzyk P, Wells TN, Hubbard RE.

Author information:

(1)BioInformatics, Glaxo Institute for Molecular Biology, Geneva, Switzerland.

Molecular modelling of the transmembrane helices of G-protein coupled receptors

is an increasingly used method to identify the possible three-dimensional

environment of key residues. Thereby site-directed mutagenesis experiments, aimed

at the understanding of the receptor-ligand interactions, can be designed in a

rational way. The modelling methods are however not generally available to

experimentalists, and often require expensive software and hardware. To overcome

these limitations, we have constructed a World Wide Web server for the automated

protein modelling of user-defined transmembrane helices. The service is freely

available at this address: http:/(/)expasy.hcuge.ch/swissmod/SWISS-MODEL.++

+html.

PMID: 9014239 [Indexed for MEDLINE]

3934. Stud Health Technol Inform. 1996;29:242-9.

Remote access to neurosurgical ICU physiological data using the World Wide Web.

Nenov V(1), Klopp J.

Author information:

(1)Division of Neurosurgery, University of California at Los Angeles School of

Medicine 90024, USA.

There is a significant demand by physicians and clinical researchers for remote

access to continuously acquired physiological patient data. Until recently such

access was technically unfeasible. However, with the recent development of

Internet-based World Wide Web (WWW) client/server applications and underlying

communication protocols, there is now a real possibility for the development of

cost-effective, platform independent solutions to this problem. We have devised a

way using existing WWW tools and minimal startup costs to provide access to

current as well as previously acquired physiological patient data. Physicians and

clinical researchers can obtain access to these data through personal computers

located in the office, at home or even through portable computers while traveling

to conferences or while on vacation.

PMID: 10163756 [Indexed for MEDLINE]

3935. Acad Radiol. 1995 Dec;2(12):1052-5.

1995 Joseph E. Whitley, MD, Award. A World Wide Web gateway to the radiologic

learning file.

Channin DS(1).

Author information:

(1)Department of Radiology, Pennsylvania State University College of Medicine,

Hershey 17033, USA.

RATIONALE AND OBJECTIVES: Computer networks in general, and the Internet

specifically, are changing the way information is manipulated in the world at

large and in radiology. The goal of this project was to develop a computer system

in which images from the Radiologic Learning File, available previously only via

a single-user laser disc, are made available over a generic, high-availability

computer network to many potential users simultaneously.

METHODS: Using a networked workstation in our laboratory and freely available

distributed hypertext software, we established a World Wide Web (WWW) information

server for radiology. Images from the Radiologic Learning File are requested

through the WWW client software, digitized from a single laser disc containing

the entire teaching file and then transmitted over the network to the client. The

text accompanying each image is incorporated into the transmitted document.

RESULTS: The Radiologic Learning File is now on-line, and requests to view the

cases result in the delivery of the text and images. Image digitization via a

frame grabber takes 1/30th of a second. Conversion of the image to a standard

computer graphic format takes 45-60 sec. Text and image transmission speed on a

local area network varies between 200 and 400 kilobytes (KB) per second depending

on the network load.

CONCLUSION: We have made images from a laser disc of the Radiologic Learning File

available through an Internet-based hypertext server. The images previously

available through a single-user system located in a remote section of our

department are now ubiquitously available throughout our department via the

department's computer network. We have thus converted a single-user, limited

functionality system into a multiuser, widely available resource.

PMID: 9419681 [Indexed for MEDLINE]

3936. Biotechniques. 1995 Dec;19(6):966-70.

Technical report. Using a world wide web server as a local organizer for protein

and DNA sequences.

Atwell R(1), Gibbins F, Upton C.

Author information:

(1)University of Victoria, British Columbia, Canada.

We have used a local World Wide Web (WWW) server to organize protein and DNA

sequences that are used frequently in our laboratory. WWW server programs are

available for most computer platforms and are easily set up with minimal computer

skills. This approach allows for the easy retrieval of sequence data, which can

then be used as input for other analysis programs. This format is especially

simple to use in conjunction with WWW database searches. The sequence files may

be served to the "public" Internet or kept private by requiring a password for

access. Other advantages are (i) sequences can be accessed from multiple computer

platforms using the appropriate WWW-browser; (ii) files can be accessed remotely

from any computer on the Internet; (iii) only a single sequence format is used,

simplifying the updating and archiving of data; and (iv) links to remote files

can also be served in addition to local files.

PMID: 8747663 [Indexed for MEDLINE]

3937. Skin Res Technol. 1995 Nov;1(4):192-9. doi: 10.1111/j.1600-0846.1995.tb00043.x.

Storage and retrieval of digital images in dermatology.

Bittorf A(1), Krejci-Papa NC(1), Diepgen TL(1).

Author information:

(1)Department of Dermatology, School of Medicine, University of Erlangen,

Germany.

BACKGROUND: Differential diagnosis in dermatology relies on the interpretation of

visual information in the form of clinical and histopathological images. Up until

now, reference images have had to be retrieved from textbooks and/or appropriate

journals. To overcome inherent limitations of those storage media with respect to

the number of images stored, display, and search parameters available, we

designed a computer-based database of digitized dermatologic images.

METHODS: Images were taken from the photo archive of the Dermatological Clinic of

the University of Erlangen. A database was designed using the Entity-Relationship

approach. It was implemented on a PC-Windows platform using MS Access\* and MS

Visual Basic®. As WWW-server a Sparc 10 workstation was used with the CERN

Hypertext-Transfer-Protocol-Daemon (httpd) 3.0 pre 6 software running.

RESULTS: For compressed storage on a hard drive, a quality factor of 60 allowed

on-screen differential diagnosis and corresponded to a compression factor of 1:35

for clinical images and 1:40 for histopathological images. Hierarchical keys of

clinical or histopathological criteria permitted multi-criteria searches. A

script using the Common Gateway Interface (CGI) enabled remote search and image

retrieval via the World-Wide-Web (W3).

CONCLUSIONS: A dermatologic image database, featurig clinical and

histopathological images was constructed which allows for multi-parameter

searches and world-wide remote access.

DOI: 10.1111/j.1600-0846.1995.tb00043.x

PMID: 27326722

3938. J Mol Graph. 1995 Oct;13(5):268-70.

Chemical collaboratories using World-Wide Web servers and EyeChem-based viewers.

Casher O(1), Rzepa HS.

Author information:

(1)Department of Chemistry, Imperial College, London, England.

We present a "proof-of-concept" model of an Internet-based chemical

collaboratory. This is based on an integration of a World-Wide Web server running

the HTTP protocol, hypertext-markup language-based browsers, molecular

visualizers based on Explorer EyeChem modules, and browsers implementing the

virtual-reality modelling language 3D (VRML) scene description.

PMID: 8603054 [Indexed for MEDLINE]

3939. Biokhimiia. 1995 Aug;60(8):1221-30.

[GeneBee-NET: An Internet based server for biopolymer structure analysis].

[Article in Russian]

Brodskiĭ LI, Ivanov VV, Kalaĭdzidis IaL, Leontovich AM, Nikolaev VK, Feranchuk

SI, Drachev VA.

A network server providing biopolymer structure databank retrieval as well as

some other biocomputing procedures for Internet users is described. Its basic

procedures consist in looking for sequence and 3D homologies (similarities).

Found homologies are used for constructing multiple alignment, for predicting RNA

and protein secondary structures as well as for constructing phylogenetic trees.

Alongside traditional methods of sequence homology retrieval, a "matrix-free"

(correlation) method is proposed. A similar procedure is used to locate protein

3D similarities. For novel procedures algorithm ideas and their possible

applications are discussed. The service ideology is based on the interaction of

server and client programs. The client program (GeneBee for IBM PC) can be used

to form queries to the server as well as to manipulate a treatment result. In the

absence of the client program the interaction with the server can be in the text

mode. The E-mail and WWW addresses for the server are as follows:

SERVE/INDY.GENEBEE.MSU.SU and WWW.GENEBEE.MSU.SU.

PMID: 7578577 [Indexed for MEDLINE]

3940. Electrophoresis. 1995 Jul;16(7):1170-4.

A Saccharomyces cerevisiae Internet protein resource now available.

Latter GI(1), Boutell T, Monardo PJ, Kobayashi R, Futcher B, Mclaughlin CS,

Garrels JI.

Author information:

(1)Cold Spring Harbor Laboratory, NY 11724, USA.

The QUEST Protein Database Center is now making available two Saccharomyces

cerevisiae protein databases via the Internet. The yeast electrophoretic protein

database (YEPD) is a database of approximately one hundred protein

identifications on two-dimensional gels. The yeast protein database (YPD) is a

database of gene names and properties of over 3500 yeast proteins of known

sequence. These databases can be accessed via a World-Wide Web (WWW) server (URL

http:@siva.cshl.org). YPD is available via public ftp (isis.cshl.org) as well, in

a spreadsheet format, and in ASCII format. When accessed via WWW, both of these

databases have hypertext links to other biological data, such as the SWISS-PROT

protein sequence database and the Saccharomyces Genome Database (SacchDB), and to

each other.

PMID: 7498160 [Indexed for MEDLINE]

3941. Electrophoresis. 1995 Jul;16(7):1131-51.

Inside SWISS-2DPAGE database.

Sanchez JC(1), Appel RD, Golaz O, Pasquali C, Ravier F, Bairoch A, Hochstrasser

DF.

Author information:

(1)Clinical Chemistry Laboratory, University Hospital, Geneva, Switzerland.

Several two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) databases

have been established and updated for more than 15 years. Only recently have

developments of computer networks and high-speed transfer protocols provided the

required tools for sharing comprehensive and hypermedia 2-D PAGE databases. This

publication describes the SWISS-2DPAGE database structure. Proteins present in

samples of human tissue, cells, cell lines and body fluids are assembled and

described in an accessible uniform format. SWISS-2DPAGE can be freely accessed

through the World-Wide Web (WWW) network on the ExPASy molecular biology server.

PMID: 7498157 [Indexed for MEDLINE]

3942. J Am Med Inform Assoc. 1995 Mar-Apr;2(2):94-101.

Managing the electronic NIH-guide for grants and contracts.

Smith PR(1), Gottesman S, Jones WK.

Author information:

(1)Department of Cell Biology, NYU Medical Center, NY 10016, USA.

This article describes the implementation of a suite of computer programs to

manage and provide access to a database containing the electronic documents that

constitute the NIH-Guide that is distributed by the NIH on a weekly basis. The

software consists of a management program that reads, processes, and stores the

incoming documents and performs erratum updates on existing documents; an

alerting program that sends selected information to users who have registered

their information needs; a viewer that can be used on the local computer to read

these documents; and a World-Wide-Web (WWW) server that can distribute the guide

to computers that run WWW client software. The design of the documentation

annotations, the management software, and the WWW server are expected to

constitute valuable models for similar projects in the future.

PMCID: PMC116243

PMID: 7743321 [Indexed for MEDLINE]

3943. J Am Med Inform Assoc. 1995 Mar-Apr;2(2):102-15.

Lessons learned from a pilot implementation of the UMLS information sources map.

Miller PL(1), Frawley SJ, Wright L, Roderer NK, Powsner SM.

Author information:

(1)Center for Medical Informatics, Yale University School of Medicine, New Haven,

CT 06510, USA.

OBJECTIVE: To explore the software design issues involved in implementing an

operational information sources map (ISM) knowledge base (KB) and system of

navigational tools that can help medical users access network-based information

sources relevant to a biomedical question.

DESIGN: A pilot biomedical ISM KB and associated client-server software

(ISM/Explorer) have been developed to help students, clinicians, researchers, and

staff access network-based information sources, as part of the National Library

of Medicine's (NLM) multi-institutional Unified Medical Language System (UMLS)

project. The system allows the user to specify and constrain a search for a

biomedical question of interest. The system then returns a list of sources

matching the search. At this point the user may request 1) further information

about a source, 2) that the list of sources be regrouped by different criteria to

allow the user to get a better overall appreciation of the set of retrieved

sources as a whole, or 3) automatic connection to a source.

RESULTS: The pilot system operates in client-server mode and currently contains

coded information for 121 sources. It is in routine use from approximately 40

workstations at the Yale School of Medicine. The lessons that have been learned

are that: 1) it is important to make access to different versions of a source as

seamless as possible, 2) achieving seamless, cross-platform access to

heterogeneous sources is difficult, 3) significant differences exist between

coding the subject content of an electronic information resource versus that of

an article or a book, 4) customizing the ISM to multiple institutions entails

significant complexities, and 5) there are many design trade-offs between

specifying searches and viewing sets of retrieved sources that must be taken into

consideration.

CONCLUSION: An ISM KB and navigational tools have been constructed. In the

process, much has been learned about the complexities of development and

evaluation in this new environment, which are different from those for Gopher,

wide area information servers (WAIS), World-Wide-Web (WWW), and MOSAIC resources.

PMCID: PMC116244

PMID: 7743314 [Indexed for MEDLINE]

3944. AJR Am J Roentgenol. 1995 Feb;164(2):479-83.

A World-Wide Web radiology teaching file server on the Internet.

Richardson ML(1).

Author information:

(1)Department of Radiology, University of Washington, Seattle 98195.

Comment in

AJR Am J Roentgenol. 1995 Feb;164(2):489-91.

OBJECTIVE: Radiology departments have traditionally used film-based collections

of interesting cases for teaching purposes. Film-based files are expensive to

create and duplicate, and they physically occupy considerable space. As one

solution to these problems, our department created an on-line radiology teaching

file in digital format.

MATERIALS AND METHODS: Our teaching file resides on a Macintosh Quadra 700

computer that is connected to the Internet, a worldwide network of interconnected

computers, via our campus Ethernet network. Our digital teaching file images and

text are composed in HyperText Markup Language (HTML) and are made available to

the world with Webserver software known as MacHTTP. These teaching files are

accessed using World-Wide Web (WWW) client software such as Mosaic, MacWeb, or

Netscape.

RESULTS: Our digital teaching file is available at no charge to anyone in the

world with access to the Internet and WWW client software. Our radiology

residents can access this file via several workstations in our department. Mosaic

is an easy-to-use interface, and the use of our digital teaching file has

increased significantly. In the 3 months since its creation, our teaching file

has been accessed not only by our radiology residents but also by hundreds of

other users in 33 countries.

CONCLUSION: Use of Mosaic and the WWW format has resulted in an easy-to-use

hypertext interface to the Internet, which allows even persons with little

computer experience to navigate through the Internet, read text files, view

images (stills and movies), and download files by merely pointing with the mouse

and clicking on items of interest. This has allowed us to maintain a central

teaching file that is physically small and easy to share with all the hospitals

in our system. We invite the worldwide radiology community to access these files

and to submit cases from their own teaching files to share with the rest of the

world.

DOI: 10.2214/ajr.164.2.7839993

PMID: 7839993 [Indexed for MEDLINE]

3945. Appl Theor Electrophor. 1995;5(2):55-72.

The Protein Disease Database of human body fluids: II. Computer methods and data

issues.

Lemkin PF(1), Orr GA, Goldstein MP, Creed GJ, Myrick JE, Merril CR.

Author information:

(1)Image Processing Section/LMMB, NCI-FCRDC/NIH, Frederick, MD 21702, USA.

The Protein Disease Database (PDD) is a relational database of proteins and

diseases. With this database it is possible to screen for quantitative protein

abnormalities associated with disease states. These quantitative relationships

use data drawn from the peer-reviewed biomedical literature. Assays may also

include those observed in high-resolution electrophoretic gels that offer the

potential to quantitate many proteins in a single test as well as data gathered

by enzymatic or immunologic assays. We are using the Internet World Wide Web

(WWW) and the Web browser paradigm as an access method for wide distribution and

querying of the Protein Disease Database. The WWW hypertext transfer protocol and

its Common Gateway Interface make it possible to build powerful graphical user

interfaces that can support easy-to-use data retrieval using query specification

forms or images. The details of these interactions are totally transparent to the

users of these forms. Using a client-server SQL relational database, user query

access, initial data entry and database maintenance are all performed over the

Internet with a Web browser. We discuss the underlying design issues, mapping

mechanisms and assumptions that we used in constructing the system, data entry,

access to the database server, security, and synthesis of derived two-dimensional

gel image maps and hypertext documents resulting from SQL database searches.

PMID: 8573600 [Indexed for MEDLINE]

3946. Medinfo. 1995;8 Pt 2:1528.

Delivering health information databases on World Wide Web at the National

University of Singapore.

Lun KC(1), Tan TW, Gopalakrishnakone P, Loh S.

Author information:

(1)Department of Community, Occupational, and Family Medicine, National

University of Singapore.

The National University of Singapore (NUS) is one of the first medical schools in

Asia to exploit the use of the World Wide Web on the Internet for the delivery of

health information databases. Its WWW server was established in 1993 by the NUS

Biocomputing Research and User Support (BRUS) technology group in collaboration

with the Computer Resource Planning committee of the Faculty of Medicine. As a

result of the early recognition of the powerful platform on which health

information services can be delivered worldwide, the NUS effort has been

accredited with a number of Internet firsts in the area of health informatics.

The following are some of the NUS achievements: NUS-NCI CancerNet on the Web. The

NUS developed and implemented the first WWW version of the popular CancerNet

database offered by the National Cancer Institute, NIH, USA. Health Info-Com

Network Medical Newsletter. The NUS developed and implemented the first WWW

version of the medical newsletter, MEDNEWS which is edited by Dr. David Dodell,

USA. It is now mirrored by the University of Pennsylvania in the United States

and De Montfort University, U.K. Poisons Information Database. This first WWW

implementation of a database on known plant, snake and other animal toxins with

directories of antivenoms, toxinologists and poisons control centers around the

world is offered by the NUS Venom and Toxin Research Group. HistoNet. This is a

large collection of histology specimens from the NUS Department of Anatomy.

MEDISTAT. This is the first WWW implementation of a Health and Population

Statistical Database which contains information for Singapore, selected Asian

countries and aggregate data for world regions. The Singapore Biotechnology

Database. This database features companies and organizations involved in

biotechnology and related activities in Singapore. Efforts are continuing to

offer more value-added health information databases on the NUS WWW server and to

link the server with other top-class information centers worldwide. Our mission

is to identify the National University of Singapore as a global health

information hub on the Internet.

PMID: 8591490 [Indexed for MEDLINE]

3947. Proc Int Conf Intell Syst Mol Biol. 1995;3:367-75.

Identification of human gene structure using linear discriminant functions and

dynamic programming.

Solovyev VV(1), Salamov AA, Lawrence CB.

Author information:

(1)Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030,

USA.

Development of advanced technique to identify gene structure is one of the main

challenges of the Human Genome Project. Discriminant analysis was applied to the

construction of recognition functions for various components of gene structure.

Linear discriminant functions for splice sites, 5'-coding, internal exon, and

3'-coding region recognition have been developed. A gene structure prediction

system FGENE has been developed based on the exon recognition functions. We

compute a graph of mutual compatibility of different exons and present a gene

structure models as paths of this directed acyclic graph. For an optimal model

selection we apply a variant of dynamic programming algorithm to search for the

path in the graph with the maximal value of the corresponding discriminant

functions. Prediction by FGENE for 185 complete human gene sequences has 81%

exact exon recognition accuracy and 91% accuracy at the level of individual exon

nucleotides with the correlation coefficient (C) equals 0.90. Testing FGENE on 35

genes not used in the development of discriminant functions shows 71% accuracy of

exact exon prediction and 89% at the nucleotide level (C = 0.86). FGENE compares

very favorably with the other programs currently used to predict protein-coding

regions. Analysis of uncharacterized human sequences based on our methods for

splice site (HSPL, RNASPL), internal exons (HEXON), all type of exons (FEXH) and

human (FGENEH) and bacterial (CDSB) gene structure prediction and recognition of

human and bacterial sequences (HBR) (to test a library for E. coli contamination)

is available through the University of Houston, Weizmann Institute of Science

network server and a WWW page of the Human Genome Center at Baylor College of

Medicine.

PMID: 7584460 [Indexed for MEDLINE]

3948. Yearb Med Inform. 1995;(1):86-97.

DIOGENE 2, a distributed Hospital Information System with an emphasis on ist

Medical Information Contenct.

Scherrer JR, Lovis C, Borst F.

DIOGENE 1 has been a mainframe-based centralised HIS with a star network of

communication operating on a daily basis with 120 nursing ward units since 1978.

Together the limited and costly growth capabilities of such a system with its

extreme difficulty in cooperating jointly with other heterogeneous medical

systems, with the need for faster networking expansions, led to the new design of

a distributed architecture called DIOGENE 2. In 1989, a migration process between

DIOGENE 1 and DIOGENE 2 was initiated and is now on the verge of being achieved.

During the time of this new expansion of the HIS, it has been easy to cooperate

with the decentralisation process of the new hospital organisation as well as

facilitating the integration of new functionalities like i.e. new WIS

architecture, medical office patient histories, integration based upon PCs with

UNIX based client/server platforms. That approach combines the handling of

paragraphs structured patient records with the use of medical natural language

processing and semi-automatic encoding as well. Amongst these new functionalities

the PACS are associated with image manipulation platforms called OSIRIS for X-Ray

images as well as other tools devoted to molecular biology and genetics up to the

ExPASy server on Internet using WWW / Mosaic which is accessible from all over

the world. The distributed architecture appears well suited not only for the

integration of these new functionalities but to keep them growing as smoothly as

possible.

PMID: 27668773

3949. Nucleic Acids Res. 1994 Sep;22(17):3610-5.

The SBASE protein domain library, release 3.0: a collection of annotated protein

sequence segments.

Pongor S(1), Hátsági Z, Degtyarenko K, Fábián P, Skerl V, Hegyi H, Murvai J,

Bevilacqua V.

Author information:

(1)International Centre for Genetic Engineering and Biotechnology, Trieste,

Italy.

SBASE 3.0 is the third release of SBASE, a collection of annotated protein domain

sequences. SBASE entries represent various structural, functional, ligand-binding

and topogenic segments of proteins as defined by their publishing authors. SBASE

can be used for establishing domain homologies using different database-search

tools such as FASTA [Lipman and Pearson (1985) Science, 227, 1436-1441], and

BLAST3 [Altschul and Lipman (1990) Proc. Natl. Acad. Sci. USA, 87, 5509-5513]

which is especially useful in the case of loosely defined domain types for which

efficient consensus patterns can not be established. The present release contains

41,749 entries provided with standardized names and cross-referenced to the major

protein and nucleic acid databanks as well as to the PROSITE catalogue of protein

sequence patterns. The entries are clustered into 2285 groups using the BLAST

algorithm for computing similarity measures. SBASE 3.0 is freely available on

request to the authors or by anonymous 'ftp' file transfer from <

ftp.icgeb.trieste.it >. Individual records can be retrieved with the gopher

server at < icgeb.trieste.it > and with a www-server at <

http:@www.icgeb.trieste.it >. Automated searching of SBASE by BLAST can be

carried out with the electronic mail server < sbase@icgeb.trieste.it >. Another

mail server < domain@hubi.abc.hu > assigns SBASE domain homologies on the basis

of SWISS-PROT searches. A comparison of pertinent search strategies is presented.

PMCID: PMC308330

PMID: 7937068 [Indexed for MEDLINE]

3950. Nucleic Acids Res. 1994 Sep;22(17):3502-7.

Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994.

Gutell RR(1).

Author information:

(1)MCB Biology, University of Colorado, Boulder 80309-0347.

A collection of diverse 16S and 16S-like rRNA secondary structure diagrams are

available. This set of rRNAs contains representative structures from all of the

major phylogenetic groupings--Archaea, (eu)Bacteria, and the nucleus,

mitochondrion, and chloroplast of Eucarya. Within this broad phylogenetic

sampling are examples of the major forms of structural diversity currently known

for this class of rRNAs. These structure diagrams are available online through

our computer-network WWW server and anonymous ftp, as well as from the author in

hardcopy format.

PMCID: PMC308311

PMID: 7524024 [Indexed for MEDLINE]

3951. Mamm Genome. 1994 Jun;5(6):372-5.

The Portable Dictionary of the Mouse Genome: a personal database for gene mapping

and molecular biology.

Williams RW(1).

Author information:

(1)University of Tennessee, Health Science Center, Department of Anatomy and

Neurobiology, Memphis 38163.

The Portable Dictionary of the Mouse Genome is a database for personal computers

that contains information on approximately 10,000 loci in the mouse, along with

data on homologs in several other mammalian species, including human, rat, cat,

cow, and pig. Key features of the dictionary are its compact size, its network

independence, and the ability to convert the entire dictionary to a wide variety

of common application programs. Another significant feature is the integration of

DNA sequence accession data. Loci in the dictionary can be rapidly resorted by

chromosomal position, by type, by human homology, or by gene effect. The

dictionary provides an accessible, easily manipulated set of data that has many

uses--from a quick review of loci and gene nomenclature to the design of

experiments and analysis of results. The Portable Dictionary is available in

several formats suitable for conversion to different programs and computer

systems. It can be obtained on disk or from Internet Gopher servers

(mickey.utmen.edu or anat4.utmen.edu), an anonymous FTP site (nb.utmem.edu in the

directory pub/genedict), and a World Wide Web server

(http://mickey.utmem.edu/front.html).

PMID: 8043953 [Indexed for MEDLINE]

3952. Trends Biochem Sci. 1994 Jun;19(6):258-60.

A new generation of information retrieval tools for biologists: the example of

the ExPASy WWW server.

Appel RD(1), Bairoch A, Hochstrasser DF.

Author information:

(1)University Hospital of Geneva, Switzerland.

PMID: 8073505 [Indexed for MEDLINE]

3953. Proc Annu Symp Comput Appl Med Care. 1994:103-7.

OncoLink: a multimedia oncology information resource on the Internet.

Buhle EL Jr(1), Goldwein JW, Benjamin I.

Author information:

(1)University of Pennsylvania School of Medicine, Philadelphia.

This paper describes OncoLink, the first multimedia World-Wide-Web (WWW) and

gopher server focusing on cancer information for both the health care

professional and the patient. OncoLink provides an internetworked hypertext and

multimedia resource linking people, computers and information together in an easy

to use fashion. Our objective in developing OncoLink is to provide comprehensive

and timely information about many aspects of oncology for both patients and

healthcare providers. Specifically, OncoLink's purposes are: (1) the rapid

dissemination of information relevant to treatment of cancer and concomitant

problems; (2) education of health care personnel (at all levels) in the field;

(3) education of patients and families of patients who have cancer; (4) posting

of clinical trials and eligibility criteria; (5) the rapid collection and

dissemination of quality, peer-reviewed information pertinent to oncology in

general and specific subspecialties; (6) provide a well-organized, frequently

updated hypertext system to access other quality cancer information resources on

the Internet. OncoLink attempts to provide one-stop shopping for the patient,

healthcare provider, researcher or Internet browser searching for cancer-related

information. Since its inception on March 7, 1994, OncoLink has averaged more

than 36,000 accesses per month from around the world. While also accessible by

text-based gopher servers, preliminary observations infer increased use of

multimedia and hypertext documents over traditional text-only resources. From the

large following of users, it is clear that electronic dissemination of high

quality, peer-reviewed cancer information is very popular. We conclude OncoLink

is both useful and has wide interest in the international community.(ABSTRACT

TRUNCATED AT 250 WORDS)

PMCID: PMC2247930

PMID: 7949867 [Indexed for MEDLINE]